Université de Sherbrooke

Exploration des méthodes de séquençage pour une identification optimale des snoRNAs

Par
Fabien Dupuis-Sandoval
Programmes de Biochimie

Mémoire présenté à la Faculté de médecine et des sciences de la santé
en vue de l'obtention du grade de maitre ès sciences (M. Sc.)
en Biochimie

Sherbrooke, Québec, Canada
Octobre, 2017

Membres du jury d'évaluation
Michelle Scott, Biochimie
Martin Bisaillon, Biochimie
Pierre-Étienne Jacques, Biologie, Université de Sherbrooke

Résumé

Exploration des méthodes de séquençage pour une identification optimale des snoRNAs

Par
Fabien Dupuis Sandoval
Programmes de Biochimie

Mémoire présenté à la Faculté de médecine et des sciences de la santé en vue de l'obtention du diplôme de maitre ès sciences (M.Sc.) en Biochimie, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Québec, Canada, J1H 5N4

Des avancées récentes dans le domaine du séquençage de prochaine génération ont ouvert une panoplie de façons de générer des données. Toutefois, chaque nouvelle méthode dévelopée est souvent appropriée à la caractérisation d'un seul type de phénomène ou de molécules. L'objectif de cette analyse est d'identifier la manière la plus appropriée de générer et traiter les données pour étudier les petits ARNs nucléolaires, snoRNAs. Récemment, ceux-ci ont été révélés comme des acteurs dans une variété de fonctions alternatives comme l'épissage alternatif, la résistance au choc oxidatif et l'état de la chromatine. Il est donc impératif de trouver une méthode qui puisse traiter une large quantité de données contenant les snoRNAs et leurs intéracteurs pour découvrir les rôles encore inexplorés des snoRNAs. Dans cette optique, un nouveau protocole a été élaboré. Cette nouvelle suite d'analyses s'appuie sur une reverse transcriptase isolée d'un intron de groupe II bactérien qui affiche une meilleure représentation des petits ARNs structurés comme les tRNAs et les snoRNAs. En effet, quand les données générées à travers la méthode de préparation des libraries pour petits ARNs standard est comparée à celle basée sur la reverse transcriptase bactérienne, cette dernière donne une meilleure représentation du compte des espèces. Ces avancées sont aussi présentes dans la méthode d'analyse informatique. La suite d'outils a été modifiée afin de permettre une meilleure détection des petits ARN non-codants. Ces modifications permettent de récupérer des millions de lectures par ensemble de données ce qui augmente le pouvoir prédictif de l'analyse.

Mots clés : petits ARNs nucléolaires, Séquençage de prochaine génération, bioinformatique, préparation de libraries de séquençage, PCR quantitatif

Summary

Exploring optimal snoRNA profiling using Next Generation Sequencing methods

By
Fabien Dupuis Sandoval
Biochemistry Program

Thesis presented to the Faculty of medicine and health sciences for the obtention of Master degree diploma maitre ès sciences (M.Sc.) in Biochemistry, Faculty of medicine and health sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada, J1H 5N4

Recent advances in Next-Generation Sequencing protocols have opened a variety of ways to generate data. However, each newly developed methodology is most suited to represent a certain phenomenon or molecule. The object of this analysis is to identify the most appropriate way to generate and process data to study the snoRNAs, or small nucleolar RNA. Recently, snoRNAs have been revealed as taking part in a variety of unexpected alternative functions such as splicing, resistance to oxidative shock and chromatin unwinding. Finding a method to generate and treat a large quantity of data containing snoRNAs and their potential interactors could highlight some of their unexplored roles within the cell. To tackle the problem, a new protocol was put forward. This new pipeline relies on a reverse transcriptase isolated from a bacterial group II intron which boasts a better representation of structured small RNAs such as tRNAs and snoRNAs. Indeed, when compared to data created by using the standard small RNA preparation protocol, the sequencing data generated through the group II intron retrotranscriptase gives a much fairer representation. These improvements are also present in the bioinformatics pipeline. The workflow was changed to facilitate the detection of ncRNAs. These modifications rescue millions of reads, further increasing the power of the analysis. Ultimately, such corrections increase the predictive power of sequencing data.

Keywords : snoRNA, Next-Generation Sequencing, bioinformatics, library preparation, qPCR

Table of Contents

LIST OF TABLES:

LIST OF FIGURES:

LIST OF ABBREVIATIONS:

Abbreviations:

RNA species

RNA	ribonucleic acid
caRNA	chromatin-associated RNA
ncRNA	non-coding RNA
mRNA	messenger RNA
miscRNA	miscellaneous RNA
snoRNA	small nucleolar RNA
miRNA	microRNA
tRNA	transfer RNA
lincRNA	long intergenic non-coding RNA
sdRNA	small nucleolar derived RNA
sno-lncRNA	Long non-coding RNA with snoRNA ends
rRNA	ribosomal RNA

Datasets

BURz	bacterial unfragmented ribodepleted
BFRz	bacterial fragmented ribodepleted
VUSs	viral unfragmented size selection
B*Rz	bacterial ribodepleted

Techniques

qPCR	quantitative polymerase chain reaction
dNTPs	deoxynucleotides triphosphate

Units

CPM	count per million
TPM	transcript per million
FPKM	fragments per kilobase of exon per million reads mapped
Ct	threshold cycle

Various

nt	nucleotide
cDNA	complementary DNA
PWS	Prader-Willi Syndrome
NGS	Next-Generation Sequencing

ACKNOWLEDGEMENTS:

I would like to direct the most generous thanks I can muster to my supervisor Pr. Michelle Scott and Pr. Sherif Abou Elela for their guidance and support through my master's. I also need to acknowledge the incredible work and energy invested in creating sequencing libraries by Sonia Couture from Pr. Abou Elela's lab.

To the members of the Plateforme RNomique Genome Quebec, I wish to direct my deepest thanks for their assistance with the qPCR data acquisition and validation. I would like to insist on the importance of the members from Lambowitz's lab, Douglas Wu and Ryan Nottingham which assisted as much as they could with the analysis. I am thankful to the Mammouth team at Sherbrooke University for their prompt assistance whenever there were issues with the nodes or a program that needed setting up.

Lastly, I would like to thank my family for tolerating my long absences and hectic behaviour through the years.

1. INTRODUCTION

1.1 General overview

1.1.1 Basic overview of snoRNAs

Small nucleolar RNAs, snoRNAs, are small non-coding RNAs (ncRNAs) species localized in the nucleolus and expressed throughout eukaryotes (Bachellerie, Cavaillé, & Hüttenhofer, 2002; Hoeppner & Poole, 2012). As small ncRNAs, snoRNAs' length ranges, most often, between 60 and 200 nucleotides (nt) in humans, but can reach up to 1000 nt in yeast (Dieci, Preti, & Montanini, 2009). To add to this wide diversity of sizes, their genomic localization can also widely vary from organism to organism. Plant snoRNAs have their own transcriptional units while human snoRNAs are found to be preferentially, over 90% of them, encoded within introns, often of coding genes related to their functions (Dieci et al., 2009). The snoRNA presence within introns affects their biogenesis. Intron encoded snoRNAs are transcribed with their host simultaneously (Tycowski, Shu, & Steitz, 1993). The introns are excised from the pre-messenger RNA (pre-mRNA) into a lariat which the human debranching enzyme, hDBR1, linearizes (Petfalski, Dandekar, Henry, & Tollervey, 1998). At this point, the core proteins of the ribonucleoprotein (RNP) complex, are already bound to the snoRNA (Ballarino, Morlando, Pagano, Fatica, & Bozzoni, 2005). Most often this complex has a single associated function, referred to as canonical, the maturation of ribosomal RNAs (rRNAs). In that function, snoRNAs act as guide sequences that match by complementarity to target sequences found on the rRNAs. At this point, the ribonucleoprotein (snoRNP) complex bound to the snoRNA modifies chemically bases on the rRNA (Smith & Steitz, 1997). The type of modification is based on the protein complement attached to the snoRNA which, in turns, is dependent on the snoRNA family. The two families, H/ACA box snoRNAs and C/D box snoRNAs, are named after the conserved sequence elements, or boxes, found in each family.

1.1.2 H/ACA snoRNA structural features

The H/ACA box family is thus named for the presence of a H box (ANANNA, N being any nucleotide) and an ACA box. The guide regions, responsible for pairing with the target, are bipartite bulges of 10 to 20 nt distributed in the hairpins (Figure 1). The hairpins exhibit a high prevalence in base pairing making the H/ACA box snoRNAs a highly structured RNA family. The ACA box is highly conserved and located 3 nt from the 3' end (Ganot, Caizergues-ferrer, and Kiss 1997; Reichow et al. 2007). This family accounts for the longer snoRNAs, with most of the members being between 120 and 160 nt, in terms of length.

1.1.3 H/ACA snoRNP biogenesis

The formation of the mature H/ACA snoRNP complex involves the assembly of a protein complex (Figure 1) composed of dyskerin (NAP57), SHQ1 and Naf1 (Hong Li, 2008; Li et al., 2011; Walbott et al., 2011). Dyskerin is the main catalytic unit of the complex, responsible for pseudouridylation of the target RNA species. SHQ1 binding to NAP57 prevents non specific RNA binding events and their improper linkage might be responsible for Dyskeratosis Congenita, a rare, congenital, progressive bone marrow failure disorder (Grozdanov, Fernandez-Fuentes, Fiser, & Meier, 2009). The Naf1, on the other hand, prevents an immature complex from having any activity. However, Naf1 is swiftly replaced by Gar1 to yield the active, mature ribonucleoprotein complex (RNP) (S Li et al., 2011). The RNP's dyskerin binds to the snoRNA. A digestion by exonucleases of the 5' and 3' ends of the snoRNA template leaves the mature snoRNP. Once processed by the exonucleases, the overall structure adopted by H/ACA snoRNAs is that of a stem-hinge-stem with the H box located in between both stems (Bachellerie, Cavaillé, & Hüttenhofer, 2002). The H box has been shown in recent studies to be dyskerin's preferred binding site (Kishore et al., 2013).

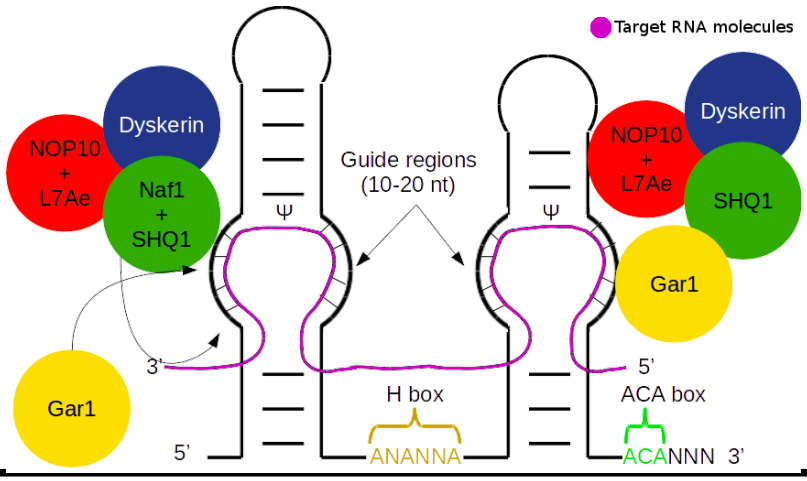


Figure 1: H/ACA box snoRNAs structural elements and their associated core proteins. H/ACA box snoRNAs have two conserved boxes, H box (ANANNA) and ACA box (ACA). Two guide regions pair up with their target RNA molecules. The assembly of a mature H/ACA snoRNA requires a full complement of proteins here depicted are NOP10, L7Ae, Dyskerin, Naf1, SHQ1 and Gar1.

1.1.4 C/D snoRNA structural features

The second family, the C/D box snoRNAs, accounts for the most genes being close to a 5:2 ratio to the other family (Lestrade & Weber, 2006). The C/D box snoRNAs family is comprised of the smaller species, often ranging between 70 and 120 nt (Scott et al., 2012). The conserved sequence elements are the C and D boxes, however, unlike H/ACA snoRNAs, there are also often duplicates of theses sequences, though often degenerate, called C' and D' (Jorjani et al., 2016; Kiss-Laszlo, Henry, & Kiss, 1998; Samarsky, Fournier, Singer, & Bertrand, 1998). The C box sequence is RUGAUGA (R being a purine) while the D box sequence is CUGA. The C and D boxes are located toward the 5' and 3' ends of the molecule, respectively, whereas the C' and D' are closer to the middle (Figure 2). The guide regions are stretches of 10-21 nt, like in the H/ACA snoRNAs, however they do not share in the bipartite nature of the H/ACA snoRNAs guide regions. Both regions are found directly upstream from the D and D' boxes (Cavaillé & Bachellerie, 1998). Further upstream, at the pairing between the C box and D box, there is the kink-turn, also labelled k-turn, which is created from two consecutive G-A pairings. This non Watson-Crick pairing results in a sharp turn in the RNA's structure (Henras, Dez & Henry, 2004).

1.1.5 C/D snoRNP biogenesis

The C/D snoRNP biogenesis is similar in many ways to the step-wise assembly of the H/ACA box snoRNAs (Figure 2). The first step is the formation of a protein complex. The 15.5K protein, analogous to the snu13p found in yeast, forms a complex with NOP58 (Bizarro et al., 2014). The 15.5K protein starts the folding by binding the C and D boxes which causes the creation of the k-turn (Watkins et al., 2000; Watkins, Dickmanns, & Luhrmann, 2002). NUFIP, analogous to the yeast's rsa1p, is then bound to the 15.5K protein and prevents the complex from carrying out its activity, akin to Gar1 for the H/ACA snoRNAs. NUFIP is also responsible for enhancing the binding of the complex to the snoRNA component, and the recruitment of the chaperone HSP90-R2TP (Boulon, Bertrand, & Pradet-Balade, 2012; McKeegan, Debieux, Boulon, Bertrand, & Watkins, 2007; Rothé et al., 2017). Fibrillarin and then, NOP56 sequentially bind to the complex and release the NUFIP factor effectively yielding the mature C/D box snoRNP. The previously mentioned k-turn is essential as it serves as an assembly and recruitment point for the fibrillarin complex (Henras et al., 2004).

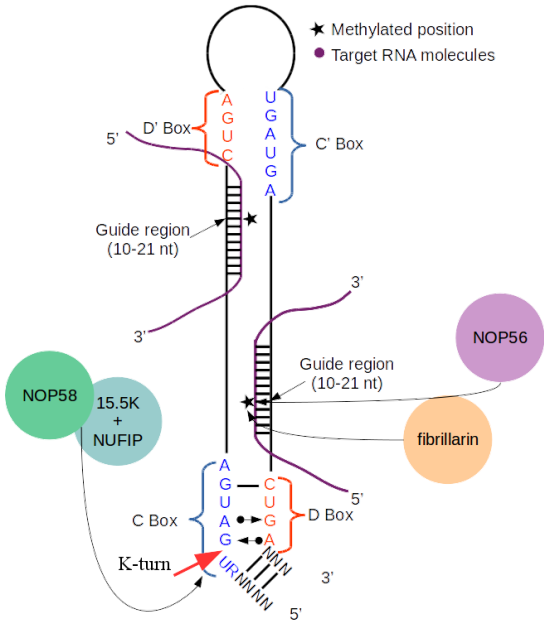


Figure 2: C/D box snoRNAs structural elements and their associated core proteins. C/D box snoRNAs have two conserved boxes, C box (RUGAUGA) and D box (CUGA). Two guide regions pair up with their target RNA molecules. A k-turn is located upstream of the C box. The assembly of a mature C/D snoRNA requires a full complement of proteins here depicted are NOP56, NOP58, fibrillarin, 15.5K and NUFIP.

1.1.6 snoRNA canonical function

The function that is most often associated with snoRNAs is the maturation of other RNA species, most often rRNA, but also tRNAs, snRNAs and other snoRNAs, through chemical modifications (Kawaji et al., 2008; Kishore et al., 2013). As previously mentioned, the type of modification depends on the core proteins associated to the snoRNPs which, is itself dependent on the family of snoRNA. The H/ACA box family is responsible for the pseudouridylation of its targets through the protein dyskerin. The guide regions are located on both stems. This, in turns, explains why the pseudouridylation occurs 14-15 nt upstream from either the H or ACA box (Ganot, Bortolin, and Kiss 1997). The modification is an isomerization of the uridine base. The most noteworthy property of pseudouridine, compared to other bases, is the presence of the hydrogen bond donor (Ge & Yu, 2013). On the other hand, the C/D box family is associated to its fibrillarin's methylation activity. The methylation occurs on the 2' hydroxyl group of the ribose (Cavaillé & Bachellerie, 1998; C. M. Smith & Steitz, 1997). The target sequence is bound to the snoRNA's guide regions, 5 nt upstream of both the D and D' boxes (Tycowski, Smith, Shu, & Steitz, 1996). The presence of two guide regions in C/D box snoRNAs indicate that up to 2 substrates can be recognized simultaneously (Kiss-László, Henry, & Kiss, 1998). However, not all snoRNAs fall within these neatly defined functions. A sizeable proportion, ~ 42% of snoRNAs, were reported, back in 2015, to lack a clearly defined function because of a lack of identifiable target (reviewed in Dupuis-Sandoval, Poirier, and Scott 2015). They were labelled orphan snoRNA. A recent study examined past high-throughput data and 2'-O-methyl profiles to put this number down to 17 % (Jorjani et al., 2016).

1.1.7 Orphan snoRNAs

Orphan snoRNAs are a subset of snoRNAs lacking a clearly defined target to modify through either methylation or pseudouridylation. Recent studies have assigned targets to four orphan snoRNAs through conservation of evolutionary homology in target sequences (Kehr, Bartschat, Tafer, Stadler, & Hertel, 2014) and identified modification sites that had, until then, been ignored. However, even after removing those 4 predictions, 143 snoRNAs species remain lacking a defined function. In recent years, through the advent of wide spectrum methodologies of analyses such as next-generation sequencing (NGS) technologies, photoactivable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP), and mass spectrometry, big data has been able to shed light on unexpected phenomenon with increased statistical significance (Hafner et al., 2010; Kishore et al., 2013). One such unexpected discovery was that snoRNA can be bound to proteins other than the core RNPs. An increased diversity of interactors is relevant, as it can be interpreted as a wider range of functions in which snoRNA are potential actors. The alternative snoRNA's roles put forward, at this point in time, are linked to resistance to lipotoxic shock, chromatin unwinding and, to some extent, splicing.

The first method in which snoRNAs can act unusually, is by modifying targets that are not normally attributed to them. When looking at the species bound to the core C/D box RNPs, unlikely species were identified as modified. These ncRNAs species were snoRNAs, snRNAs, tRNAs, but, perhaps more interestingly, vault RNAs, 7 SK and 7 SL (Kishore et al., 2013).

1.2 non-canonical functions

As mentioned before, the advent of NGS technologies has allowed to generate massive amount of data about expansive biological systems. These methods are unlike those of the past where molecules had to be studied individually. This phenomenon accounts for the limited number of snoRNAs with an appreciable depth of characterization. NGS technologies have also allowed to characterize RNA species found within individual cell compartments and organelles. This ability to segregate and characterize compartments negated the common perception that the cytoplasm did not contain snoRNAs (Holley et al., 2015).

1.2.1 snoRNA and stress responses

As touched in the earlier paragraph, the lack of wide spectrum analysis methods meant that entire species were not studied for decades. SNORD32A, SNORD33 and SNORD35A, are such molecules, all localized within the introns of RPL13A. They were reported in 1996, however until 2011, their implication in lipotoxicity resistance remained a mystery (Michel et al., 2011; Nicoloso, Qu, Michot, & Bachellerie, 1996). Knockdown of the host gene and, therefore, its associated snoRNAs induced a resistance to lipotoxic and oxidative shocks (Michel et al., 2011). The potential of snoRNAs as agents of metabolic stress was further demonstrated when a RNA immunoprecipitation coupled with sequencing (RIP-seq) detected an increased snoRNA representation on the protein kinase RNA-activated (PKR), a protein involved in the stress response, when a cell is exposed to metabolic stress (Youssef et al., 2015). As such, the human stress response seems intricately linked to snoRNAs.

1.2.2 snoRNA and chromatin

SnoRNAs are used for signalling in other organisms than humans. In Drosophila cells, RNA bound to Df31 is responsible for maintaining the open “unwound” state of chromatin (Schubert & Längst, 2013). RNA molecules bound to chromatin are referred to as chromatin-associated RNAs or caRNAs. caRNAs were shown to be predominantly H/ACA snoRNAs (Schubert et al., 2012).

1.2.3 snoRNA and splicing

Classically, the sites targeted by snoRNA were located on ncRNAs and pre-rRNA, however, pulldown experiments on snoRNAs' interactors show that snoRNAs bind sites found on mRNA (Gumienny et al., 2016). However, the same study demonstrated that these sites fail to be methylated. This absence of modification might allude to a mode of action that falls outside the canonical definition. Though this is a single example, other instances of snoRNAs acting on mRNA exist. The well characterized snoRNAs species SNORD116 and SNORD115 are known to be major actors in the Prader-Willi Syndrome (Kishore & Stamm, 2006; Peters, 2008; Skryabin et al., 2007). Recently, SNORD115 was put forward as modulating the expression of SNORD116. SNORD115 affects the pre-mRNA splicing of the serotonin receptor 2C on its exon Vb, though the exact method by which splicing of the transcript remains unclear (Kishore & Stamm, 2006; Kishore et al., 2010). SNORD116, on the other hand, does not appear to affect splicing, but rather its deletion causes a change in the expression level of certain mRNAs (Cavaille, 2017; Falaleeva, Surface, Shen, de la Grange, & Stamm, 2015). The SNORD116 family was also reported to form long non-coding RNAs derived from sequences between two snoRNAs being incorporated into snoRNA long ncRNAs, sno-lncRNAs. The deletion of 5-6 Mb from the locus containing sno-lncRNAs is present in PWS (McCann & Baserga, 2012). Furthering the link between splicing and snoRNAs, these sno-lncRNAs were found to bind RBFOX2 (also called RBM9), a member of the FOX family, which are regulators of alternative splicing (Yin et al., 2012; Zhang et al., 2014). The physiological effects of the sno-lncRNA deletion on the mice are consistent with PWS. The trait common to PWS are low birth weight, followed by increased body weight gain in adulthood, increased energy expenditure and hyperphagia (Qi et al., 2016). On the other end of the spectrum, smaller fragments of snoRNAs have been described in the literature (Falaleeva & Stamm, 2013; Kawaji et al., 2008; Taft et al., 2009).

Those snoRNA derived species (sdRNAs) are between 18 and 30 nt, meaning that they are closer in length to miRNAs than to the full snoRNA template (Brameier, Herwig, Reinhardt, Walter, & Gruber, 2011). Their biogenesis is, like miRNAs', dependent of Dicer, but, unlike miRNAs', it is independent of Drosha/ DGCR8. These sdRNAs are implicated, in a similar way to miRNAs, in the silencing pathways (Brameier et al., 2011; Scott & Ono, 2011). One such sdRNA comes from the processing of SCARNA45 (Ender et al., 2008). Similar cases of known miRNAs originating from snoRNA precursors include let-7g, mir-140, mir-151 and mir-16-1 (Ono et al., 2011; Scott, Avolio, Ono, Lamond, & Barton, 2009).

1.3 Categorizing and detecting snoRNAs

1.3.1 Improving on snoRNAs' characterization through a global sequencing approach

The widespread usage of NGS has only recently taken roots. In the 1990s and early 2000s, snoRNAs were studied individually and a sizable quantity of genetic material had to be extracted in order to give a fair representation. The laborious nature of the process meant that the analysis of the most abundant species was favoured. As such, SNORD3A, SNORD118 and species with special relevance to diseases such as PWS's SNORD115 and SNORD116 were studied early on (Kass, Tyc, Steitz, & Sollner-Webb, 1990; Tyc & Steitz, 1989). By the same process, low abundance and tissue-specific snoRNAs were ignored (Table 1).

snoRNAs	Type	Target on rRNA	Reference
SNORD3A	C/D box	18 S rRNA	A. Borovjagin, S. Gerbi 2004
SNORD115	C/D box	serotonin receptor 5HT-2C mRNA	S. Kishore, S. Stamm 2006
SNORD116	C/D box	Unknown	Q.Yin, L.Yang et al. 2012
SNORD118	C/D box	Unknown	B. Peculis 1997

Table 1: Characteristics of the most characterized snoRNAs. SNORD3A's target is on 18S rRNA while SNORD115's target is located on the serotonin receptor 5HT-2C exon V. SNORD116 and SNORD118 do not have any identified rRNA target.

Until a few years ago, the technical constraints of the past were reflected in the absence of big data on which a global portrait of snoRNAs could be made. Today, the relatively low cost, speed and ease at which sequencing can be performed is responsible for the growing abundance of small ncRNAs datasets. Nonetheless, even with the massive quantity of data, the snoRNAs are not well detected. In most sequencing sets, snoRNAs are not perceived at a sufficient depth to allow characterization. There is also the added issue that the ratio between both families is always skewed toward an overwhelming, >90%, representation of the C/D box snoRNAs species (Deschamps-Francoeur et al., 2014; Kishore et al., 2013). The low number of detected H/ACA box snoRNAs limits our ability to categorize and study an important portion of all snoRNAs. Furthermore, an additional piece of the puzzle is still missing to properly judge the properties of snoRNAs, a way to compare the species to their host and to the full spectrum of possible interactors.

As of late, the identification of snoRNAs' implications in biological processes as diverse as splicing or lipotoxicity resistance have opened the possibility of other hidden partnerships between interactors that until now hadn't been considered. To identify such interactions, the snoRNAs and their potential partners need to be compared to each other and with the base expression level to theorize how one modulates the expression of the other. Protein-coding genes combined with snoRNAs genes profiles would allow to infer partners to snoRNAs in their non canonical functions. As a consequence, more depth of data, enough to encompass snoRNAs and protein-coding genes, is required. To summarize, the data has to be broad enough to capture snoRNAs and their interactors while being deep enough that statistical inference remains a possibility. As such, these specifications leave only sequencing as an option for a semi-quantitative analysis. The entire RNA species within a cell represents very heterogeneous data in terms of length, base composition and structure. The sequencing protocol has to be likewise broad and permissive to capture all the fluctuations in the snoRNAs' length and abundance.

1.3.2 Introduction to sequencing:

The first step in the needed RNAseq workflow consists of a proper isolation of the RNA species chosen for quantification through a RNA extraction protocol. The method by which RNA species are extracted and selected affects the overall scope of the study. The frequently used protocols available to extract and enrich RNAs are poly(A) selection, ribo-depletion, and size selection (Figure 3) (Zhao et al., 2014).

1.3.2.1 Poly(A) selection

Poly (A) capture protocols rely on poly (T) oligomers to affix the adaptor unto the second strand template through the amplification process (Chang, Lim, Ha, & Kim, 2014; Zhao et al., 2014). This creates a bias as only species with polyA are detected which are mostly mRNAs (Cui et al., 2010). The coverage of mRNA is also subject to a bias against the 3' extremity as it often has homopolymeric repeats (Chang et al., 2014). Human snoRNAs are predominantly encoded within intronic regions and do not have polyadenylated tails (Dieci et al., 2009). As such, human snoRNAs, among other ncRNAs, have been found to be very poorly represented in datasets generated through poly(A) selection protocols (Zhao et al., 2014).

1.3.2.2 size selection

The second method, size selection, is most often used to sequence miRNAs (Head et al., 2014). This protocol isolates small RNA species based on their molecular weight through a glass fiber column purification. The column purification allows to extract species with less than 200 nt (Shingara et al. 2005). These species are then amplified through PCR cycles and sequenced. This selection, however, removes the longer protein-coding and ncRNAs transcripts (Sheng Li et al., 2014).

1.3.2.3 Ribo-depletion:

Ribosomal RNAs are often the most abundant RNA species within untreated samples. It was estimated that rRNA species can represent 90% of the detected species (O'Neil et al. 2013). Often rRNA can very effectively be removed of RNA samples through ribo-depletion (Conesa et al., 2016). Ribo-depletion relies on the hybridization of probes to rRNAs followed by precipitation on magnetic beads (O'Neil et al., 2013).

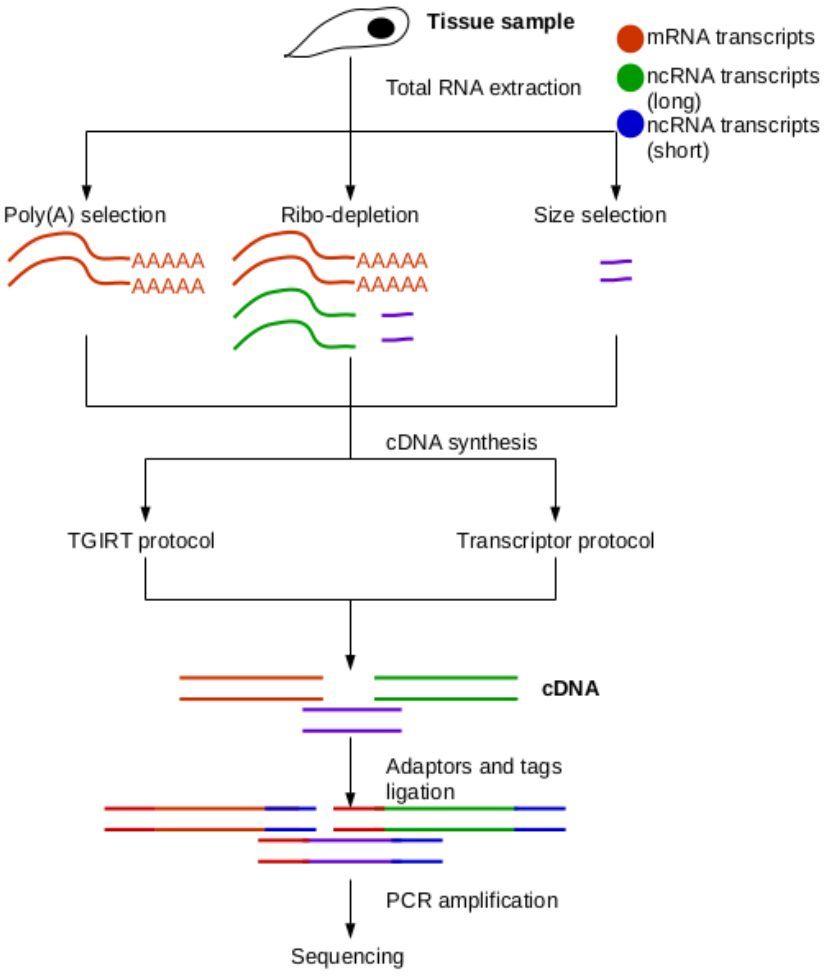


Figure 3: Sequencing workflow from RNA extraction to sequencing. The first step is to extract the total RNA from the tissue sample. The RNA is then further selected through various isolation protocols differentiating between mRNAs, longer and smaller ncRNAs. The RNA is then put through a PCR cycle which results in cDNAs. The adaptors and tags are added to the cDNA. The cDNA is then amplified either through the standard Transcriptor protocol or the TGIRT protocol before being sent for sequencing.

Extraction is usually followed by amplification of the RNA species to ensure detection by the sequencer. The result is complementary DNA species to the isolated RNA species on which two adaptor sequences and a tag were affixed through a step of adaptor ligation. These steps finalize the preparation of an appropriate library.

1.3.3 Injection of spike-ins

As an extra step, it is possible to add fixed amounts from a calibrated solution of spike-ins. These small RNAs do not have any match on the human genome. As such, they can be safely detected without the risk of being confused for another species. The spike-ins can be used as a means to estimate the presence biases in lower sizes molecules (Risso, Ngai, Speed, & Dudoit, 2014). The ERCC spike-ins are sold as 2 cocktails of molecules of varying length and base composition. ERCC spike-ins are mixed with the library samples. The concentration of each cocktail is equal and conserved variations in the perceived count after sequencing could indicate biases in the sequencing methodology (Locati et al., 2015). They can also be used to normalize species found below 100 nt, by adjusting count by a factor equal to the detected spike-ins (Nottingham et al., 2016).

The completed libraries are sent to an external site for sequencing. There, the cDNA is pooled, then bound to a flow cell by complementarity between a primer on the surface of the cell and the incorporated adaptor. Once binding to the flow cell is done, the species are amplified to create detectable clusters in a step named bridge amplification. A polymerase, nucleotides and buffer are added to the flow cell's environment to initiate the amplification process. The DNA species bridging is done by complementarity between the free adaptor sequence and a second complementary primer on the flow cell. The bridged DNA's complement sequence is transcribed by the polymerase. This process is repeated multiple times to create a cluster.

1.3.4 Sequencing-by-synthesis

The clusters are sequenced by sequentially adding bases with fluorescent dyes. The fluorescent dyes are various chromophores of specific wavelength which are freed upon binding to their complement base. These signals are intercepted by the sequencer. The sequencing apparatus attributes a quality score based on the purity and strength of the signal. This score takes the form of a Phred score, a logarithm based measure of certainty for the nature of the base sequenced. The process, as mentioned previously, is sequential, that is to say, for each base on the target sequence, bases are added until a match is found and then the following base goes through the same process. However, as bases are sequenced, the signal becomes muddled because unsuccessful reactions cause a lag in synchronization which, ultimately, accounts for a drop in the base calling confidence (Fuller et al., 2009).

1.3.4.1 Paired-end sequencing

A way to counteract the aforementioned decay in base calling confidence is to rely on paired-end sequencing. Paired-end sequencing is based on a set of complementary adaptors ligated to 3' and 5'. The presence of complementary adaptors yields both forward and reverse strands which is also done in the single-end standard protocol, however unlike the former, the reverse strand of the target sequence is conserved for further analysis.

1.3.4.2 Bypassing inherent sequencing biases

The second hurdle in generating data as unbiased as possible lies in the use of the commercially available reverse transcriptase, Transcriptor. When creating the libraries, the use of the viral reverse transcriptase has been shown to lower the representation of highly structured RNA species such as tRNAs, snoRNA and snRNAs (Nottingham et al., 2016; Zheng et al., 2015). This issue would affect our estimation of snoRNA abundance and affect every conclusion reached. An encouraged alternative to the viral reverse transcriptase was the thermostable group II intron reverse transcriptase, also known as TGIRT.

1.3.5 Group II introns

Group II introns are a family of bacterial mobile retroelements with intron-encoded reverse transcriptase and an autocatalytic RNA (Lambowitz & Zimmerly, 2004; Truong, Sidote, Russell, & Lambowitz, 2013). Group II introns have also been identified in mitochondrial's, chloroplast's, plants' and fungi's genomes. However, this family is absent of nuclear genomes (Lambowitz & Belfort, 2015). The group II introns' ability for retrohoming and retrotransposition has been linked with the appearance of spliceosomal introns, retrotransposons and telomerase. The autocatalytic component is responsible for the insertion of his cDNA into the host genome (Enyeart, Mohr, Ellington, & Lambowitz, 2014; Nottingham et al., 2016; Zheng et al., 2015). However, the most pertinent component of group II introns as far as sequencing technologies are concerned is the reverse transcriptase (Mohr et al., 2013). The RT, under normal condition, synthesizes the complement DNA (cDNA) while also having endonucleolytic capabilities (Lambowitz & Zimmerly, 2004; Nottingham et al., 2016). Structurally, the RT is similar to retroviral RTs with an extra conversed block and distal binding sites (Blocker, Mohr, Conlan, & Qi, 2005; Lambowitz & Belfort, 2015). Both the RNA and RT components can work independently from one another. This modularity, ultimately, means that for the purpose of creating sequencing libraries, the RT can be isolated from the rest of the group II intron complex (Enyeart et al., 2014). TGIRT libraries were shown to represent quite aptly tRNAs when compared to the standard illumina sequencing protocol (Zheng et al., 2015). The hypothesis of this study is that the use of the TGIRT protocol on a new generation of the illumina HiSeq sequencer would ensure a better depth and representation of snoRNAs.

1.3.6 Pipeline for the analysis of total RNAs

The analysis of the sequencing data, in itself, raises a number of problems since most methods make assumptions about the nature of the data. As such, for an analysis pipeline to be fitting, a minimum of time must be spent examining the tools available. The standard bioinformatics workflow to examine NGS data begins with an examination of the quality of the sets (Conesa et al., 2016).

Following that assessment, the data is treated to remove bases corresponding to non-genomic sequences (tags, adaptors and multiplexes) and low quality segments found within reads as the practice has been found to improve overall data quality (Shendure et al., 2008). Once the reads' quality ascertained, the process of mapping the reads to the genome through an aligner is carried out. The final step is to associate the genomic ranges found to map to reads to their corresponding genes and transcripts (Conesa et al., 2016).

1.3.7 Quality assessment

Quality assessment serves as a preliminary examination of the composition of the NGS sets. The program FastQC computes helpful metrics such as the sequence's base composition, GC content, unknown base content, duplication rates, length and quality scores (Andrews, n.d.). Different protocols to create libraries come with predicted biases. Isolation of variants of a few transcripts results in a fail of the sequence duplication analysis. The presence of adaptors within the reads fails the analysis of the Kmer content. Such scenarios can be brought up for each category. As such, implementation of checks throughout the NGS data's treatment are privileged.

The first step before the analysis of the reads is to remove all non genomic segments and reads that would, otherwise, impair proper analysis. This step incorporates multiple processes, the removal of adaptor sequences from the reads, the removal of small reads that could map at multiple locations and the trimming of the lower quality portions of the reads (Martin, 2011; Shendure et al., 2008). Removal of the adaptors is done, normally, by a flexible regular expression that matches the given sequence and returns the reads free of the adaptors.

Trimming is performed by scanning the reads from the end of the reads' sequences and removing all bases falling below a certain threshold, often a Phred score of 20 or 30. Both of the previous steps can be implemented in the same tool offering the possibility of executing both procedures simultaneously (Martin, 2011).

Once all non-genomic contamination has been removed, the reads are ready to be mapped to the genome. Multiple mapping strategies and algorithms have been implemented into stand-alone programs available to all (Bray, Pimentel, Melsted, & Pachter, 2016; Conesa et al., 2016; Dobin et al., 2013; Havgaard, Torarinsson, & Gorodkin, 2007; Kent, 2002; Kim et al., 2013; Langmead & Salzberg, 2012; Larkin et al., 2007). However, the principle remains the same. Two strings of characters are compared, one being the read, the other the reference, chromosome or genome to see where the best match for the read is located on the reference sequence. Alignment algorithms are designed to either support local or global alignment with a few programs allowing for both (Langmead & Salzberg, 2012).

Global alignment, as opposed to local alignment, requires the mapping of both ends of the supplied sequence to the target sequence. The first working global alignment method used for biological purposes used the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970). Through heuristics, this method was improved and remains used today to find the best global alignment (Kent, 2002). However, the Smith-Waterman local alignment algorithm, postulated in 1981, found popularity shortly thereafter (T. F. Smith & Waterman, 1981). These methods were pushed forward with the creation of the first sequencing-by-synthesis machines from Solexa, now owned by Illumina, and a new wave of alignment algorithms were designed. At that time, the early 2000's, tools like MAQ, BWA, SOAP and BLAT were created (Heng Li & Homer, 2010). BWA and SOAP relied on the Burrow-Wheeler Transform to improve on memory usage by compressing the index (Heng Li & Durbin, 2009). Further recent improvements on the index building and alignment principles have yielded critical advances.

1.3.8 Differences between aligners

The main differences in protocols come from the choice of alignment software and the way in which reads are annotated afterwards. Today, with the recent popularization of NGS, sequencing data has become abundant and the methods to align reads has become equally varied. The most widely used and documents alignment software are Splice Transcripts Alignment to a Reference (STAR), Tophat2, Bowtie2 and Kallisto (Conesa et al., 2016).

The first candidate, Kallisto, is a software written in python with the core aligner in C++ (Bray et al., 2016). The main quirk from Kallisto is the lack of actual alignment for each base or k-mer, rather it relies on pseudoalignment. The transcriptome is extracted from a fasta file and a De Brujin graph free of redundancy is built from k-mers as an index (Bray et al., 2016). The reads are then used as error-free paths which means mismatches between reads and transcriptome are disregarded entirely. The gene count utilizes an expectation-maximization (EM) algorithm. The reliance on a single De Brujin graph reduces the requirements in terms of CPU usage compared to other programs. Kallisto is also much faster as it does not perform complete alignments for each base. However, Kallisto cannot perform transcript discovery and extensions are hard to estimate.

Bowtie2 is an aligner based the use of full text minute (FM) index and the Burrow-Wheeler Transform. It is written in C++ (Langmead, 2013). This aligner is widely used for its ability to tolerate gap regions and perform local alignments. The biggest CPU usage from bowtie2 comes from its FM index which takes a bit more than 3 GB for the human genome. Even if the memory footprint is non negligible, any modern laptop should be able to allocate this much memory and its speed of execution is also quite respectable, being faster than the Tophat2 (Langmead & Salzberg, 2012).

The next alignment software is Tophat2, a splice junction aware program written in python and C++. Tophat2 uses bowtie2's aligner, but implements routines to take splicing of mRNA into account (Kim et al., 2013). A series of improvements have been made since the initial release of Tophat to reduce the number of core hour needed to produce the alignments. However, it still lags behind Kallisto and STAR while its alignment is quite similar to the one created by STAR.

STAR is a C++ alignment program that aimed at improving the mapping speed while allowing for new junctions discovery. It does so by looking up seeds and, in a second step, scoring them. The seed search portion relies on the Maximum Mappable Length (MMP) or mapping iteratively the longest contiguous segments of a read. The second step unites the various possibilities of seeds together and using a local alignment grading scheme assigns scores to them. The higher score is conserved as the best match and returned. STAR achieved a multiple fold faster speed of mapping than Tophat2 on simulated data (Dobin et al., 2013).

Alternatively, Cufflinks can be used to analyze RNA sequencing sets. The suite of programs Cufflinks has been created in 2009 by the Trapnell lab and remains to this day a widely used pipeline. It was designed to primarily handle sets covering the exons and protein-coding regions of the genome. This particularity is its strength since Cufflinks relies on normalization factors used specifically in protein-coding analysis, such as, FPKM (Trapnell et al, 2012). This method does not agree with comparative studies between sets because it does not normalize to negate sequencing depths differences (Conesa et al, 2016). Other normalization factors commonly used are CPM or TPM. The main difference between both method lies in TPM's adjustment for the transcript's length. Nonetheless, both TPM and CPM retain the last step of normalizing for sequencing depth. In turns, the last step allows comparison between individual counts as the sum of an experiment will always add up to a million. Ultimately, this makes CPM and TPM more relevant to a comparative study. Furthermore, the nature of our analysis, where sets are principally composed of small RNAs, made cufflinks and associated software ill-suited because it removed the possibility to account for the mobile nature of small ncRNAs. These preceding documented reasons excluded Cufflinks from any enquiry.

1.3.9 Differences between annotation strategies

Once the reads are mapped, the following step would be to assign the reads to the genes. The most popular methods are HTSeq, RSEM, bedtools multicov. The first program, RSEM, is written mainly in C++. RSEM's annotation process is anchored on an iterative fitting method called Expectation-Maximization (Kanitz et al., 2015). RSEM provides plenty of useful feedback, such as the various gene counts after normalization. The output takes the form of the gene identification, the transcripts identifications, the length, the expected count, the TPM and the FPKM.

Bedtools is a platform written in C++ for the analysis of mapped genomic data. It offers a suite of scripts useful to study the various genomic intervals returned within the BAM files (Quinlan & Hall, 2010). Multicov, one of the scripts, uses a gtf annotation file to return the number of overlaps between reads and each of the features. The information returned through this process is composed of the reference's name, start position, end position and the reads count mapping to the interval.

HTSeq is a platform for executing simple manipulations on genomic data. The platform and its associated scripts are entirely written in python. It was designed to be adapted to fit the user's needs. Within HTSeq, the script HTSeq-count provided a general template of the annotation process. The script allowed for the use of paired-end sequencing and provided ample information pertaining to its usage (Anders, Pyl, & Huber, 2015). The HTSeq output is the gene name and the associated reads count.

1.4 Objectives

As mentioned before, the snoRNAs species have, thus far, not been properly characterized because of the methods' limitations. However, with the advent of sequencing technologies, new protocols are available. Those protocols have yielded new insight into snoRNAs' alternative roles, however those studies examined proteins that were not known snoRNAs interactors and are, by extension, serendipitous events. The limited spectrum of molecules surveyed forbids a further look at the fluctuations of snoRNAs populations. As such, the first objective of our study is to find the most reliable sequencing protocol for the detection of RNAs species, specifically snoRNAs. To this end, the TGIRT protocol known for its ability to capture small, highly structured RNA species' profiles like the tRNAs will be pitted against the more conventional protocols. It is our team's hypothesis that the abundance profiles given by the TGIRT protocol will be a fairer representation of snoRNAs' abundance than any other alternative. The second objective is to construct a bioinformatics workflow able to detect qualitative and quantitative shifts in the snoRNAs. This pipeline would allow to hazard an hypothesis concerning alternative functions and interactors of snoRNAs.

2. MATERIAL AND METHODS

2.1 Generation of genomic data

2.1.1 Cell culture and transfection

The analysis and use of non simulated genomic data was invaluable to assess the presence of biases in next generation sequencing protocols. Cell types that had already been characterized would have to be used to ensure that variations from past experiments could be validated. As such, the cell type SKOV3ip1 was selected as the members of the lab personnel had experience in handling, cultivating and past datasets were readily available. SKOV3ip1 is an ovarian adenocarcinoma cell type. The cells were grown in DMEM/F12 (50/50) medium with 10% fetal bovine serum and 2mM L-glutamine. Cells were seeded at 350 000 cells per well.

2.1.2 RNA extractions

Total RNA extractions were carried out based on the protocol provided by the manufacturer, Qiagen. Following extraction, RNA integrity was confirmed by Agilent 2100 Bioanalyzer. All samples were brought to a volume of 11 µl. Random hexamers, dNTPs and RnaseOUT from Invitrogen were added, bringing the total volume of each sample to 20 µl. All samples were put through a reverse polymerase chain reaction (RT-PCR) using Transcriptor Reverse Transcriptase from Roche Diagnostics. Reverse transcription was carried out at 55 °C which is well within the optimal temperature range of 45-60 °C.

2.1.3 Library preparation and sequencing

Two batches of samples were prepared, the first in the summer of 2013, was made of 8 samples, however only 2 replicates (VUSs_1 & VUSs_2) are relevant to this analysis. While the second batch was prepared during the winter of 2016 and was made of 4 samples (BURz_1, BURz_2, BFRz_1 & BFRz_2) (Figure 4). The former batch of libraries, VUSs_1 and VUSs_2, was prepared by size selection using the mirVana toolkit by following the instructions provided by the manufacturer, ThermoFisher Scientific. Small non-coding RNA (ncRNAs) species with low molecular weight (<200 nt) were isolated.

The isolated samples were put through the TruSeq Small RNA Sample Prep kit from Illumina following the protocol provided by the manufacturer. The isolated cDNA was ligated to 3’ and 5’ adaptors included within the kit’s materials. The cDNA was reverse transcribed and amplified through PCR cycles. The reverse transcription ensured that only species with both adaptors properly ligated were amplified.

The following step, amplification, was required to generate a signal strong enough to be interpreted by the sequencer. Amplification was achieved by incubating the DNA samples in a thermal cycler through a 11 cycles PCR reaction. Verification of the samples integrity through Agilent 2100 Analyzer attested of the libraries’ purity.

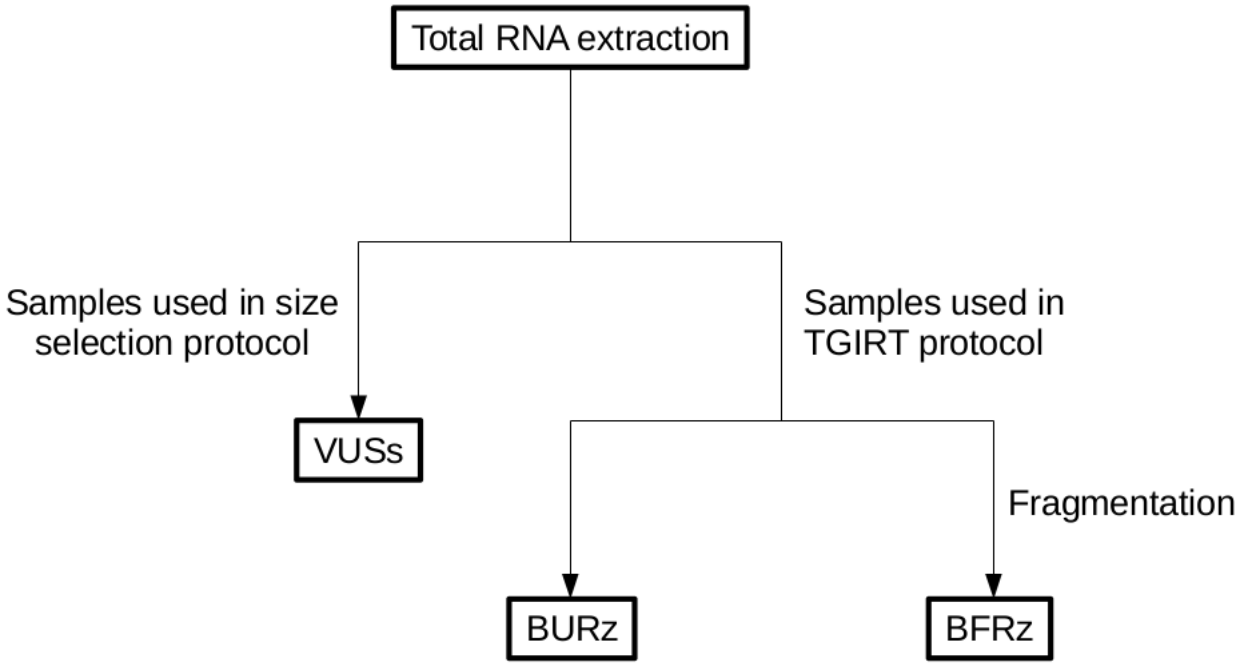


Figure 4: Divisions of total RNA samples and labelling of sequencing sets. Following a total RNA extraction, the samples were subdivided between those subjected to the library preparation for size selection protocol and those put through the TGIRT library preparation protocol. The TGIRT samples are further subdivided between those that underwent a step of fragmentation and those that did not.

The libraries, VUSs_1 and VUSs_2, were sent to McGill's and Genome Quebec Innovation center's sequencing platform to be sequenced on a HiSeq2000 sequencer from Illumina. The two samples of interest were paired-end sequenced at a read length of 100 nt and their depth are 18.5 and 16 M reads. The datasets are stored on NCBI's GEO portal under the accession number GSE55946.

The latter batch of libraries, BURz_1, BURz_2, BFRz_1 & BFRz_2, was prepared according to the specifications provided by Pr. Alan Lambowitz from the University of Texas (Nottingham et al., 2016). After each of the following steps, purification was carried out using Pr. Lambowitz's modified version of the Zymo RNA Clean & Concentrator protocol. The total RNA samples were ribodepleted using the RiboZero Gold kit from Illumina following the protocol provided by the manufacturer. The samples were mixed with a supplied spike-ins cocktail to varying ratios described by Pr. Lambowitz. The ERCC spike-ins cocktails were added, 2 µL to a 5 µL volume of the aforementioned total RNA sample (Nottingham et al., 2016). The 4 samples were subdivided. Half, BFRz_1 and BFRz_2, were fragmented using an NEBNext Magnesium RNA Fragmentation Module from New England Biolabs at a temperature of 94°C for a duration of 7 minutes. Afterwards, all of the four samples, BURz_1, BURz_2, BFRz_1 & BFRz_2, were treated with T4 polynucleotide kinase phosphatase from Epicentre.

The T4 polynucleotide kinase ensured that samples would be free of 3' phosphates and 2', 3' monophosphates which have been previously described to prevent the TGIRT-III's ability for template switching (Mohr et al., 2013) and adopted into Pr. Lambowitz's protocol. All of the 4 samples, BURz_1, BURz_2, BFRz_1 & BFRz_2, were put through reverse transcription with 1µM TGIRT-III RT from InGex, LLC and 5′ AppDNA/RNA Ligase from New England Biolabs for a duration of 15 minutes at a temperature of 60 °C and amplified for 12 cycles in a thermal cycler. Amplification was carried out as previously described (Nottingham et al., 2016). Sequencing for the 4 samples, BURz_1, BURz_2, BFRz_1 & BFRz_2, was done on site at the University of Texas's Genomic Sequencing And Analysis Facility in Houston on a Hiseq 4000 from Illumina. Reads were paired-end at 150 nt and each dataset contains approximately 30 millions reads.

2.1.4 Quantitative Polymerase Chain Reaction (qPCR)

qPCR were run in house at the Plateforme RNomique Genome Quebec found at the address (http://rnomics.med.usherbrooke.ca/). A list of protein coding and snoRNA genes was submitted to the members of the Plateforme Rnomique. Primers were designed based on the lack of sequence repetition and folding of the targeted RNA molecule (Table 2). Total RNA was extracted from SKOV3ip1 cells using TRIzol from Invitrogen with chloroform following the manufacturer's protocol. The recovered RNA was purified using the Rneasy Mini Kit column from Qiagen. A DNAse treatment was carried out as per the manufacturer's instructions. RNA integrity was confirmed by Agilent 2100 Bioanalyzer.

The extracted RNA was put through reverse transcription using 1.1 µg from the total RNA with Transcriptor reverse transcriptase, random hexamers, dNTPs from Roche Diagnostics and 10 units RNAseOUT from invitrogen. Forward and reverse primers were suspended in 20-100 µM Tris-EDTA solution from IDT and diluted as a primer pair to 1 µM in Rnase Dnase free water from IDT. The PCR reactions were carried in 10 µL in a 96 well plates on a CFX-96 thermocycler with a 5 µL volume of 2X iTag Universal SYBR Green Supermix, 3 µL of cDNA, and 2 µL from the previously mentioned primer pair solution. The thermocycler and iTag solution came from BioRad. The thermocycler was brought to 95 °C for 3 minutes and then, the cycling conditions were set as 50 cycles of 15 seconds at 95 °C, followed by 30 seconds at 60 °C, and then 30 seconds at 72 °C. Relative expression levels were assessed using the qBASE framework (Hellemans, Mortier, De Paepe, Speleman, & Vandesompele, 2007). No-template runs were carried out as negative controls (Brosseau et al., 2010).

Gene	primer forward	primer reverse	primer sequence forward	primer sequence reverse
RN7SK	RN7SK.q.F1	RN7SK.q.R1	CGGTCTTCGGTCAAGGGTATACG	AGCGCCTCATTTGGATGTGTCT
RN7SL2	RN7SL2.q.F1	RN7SL2.q.R1	AGGTCGGAAACGGAGCAGGTC	CGGGGTCTCGCTATGTTGCT
RNY1	RNY1.q.F1	RNY1.q.R1	TTGATTGTTCACAGTCAGTTACAGATCG	AGTCAAGTGCAGTAGTGAGAAGGG
RNY3	RNY3.q.F1	RNY3.q.R1	TGGTGTTTACAACTAATTGATCACAACCAG	AGCAGTGGGAGTGGAGAAGGAACAAAG
SCARNA18	SCARNA18.q.F1	SCARNA18.q.R1	TGCTTCTCATTCCTTGGGAGCA	TGTTGGAGAAATATACTACCACTCAACCT
SCARNA1	SCARNA1.q.F1	SCARNA1.q.R1	ACCGAGCTGTCTATATCCTAGCCT	ACTGGGCTTAAAAGACTCATGGCT
SCARNA22	SCARNA22.q.F1	SCARNA22.q.R1	GTCCTGACCTGTCCTCTGTGAGC	TGTACTGAAAACCGTGGTGTCCT
SNORA23	SNORA23.q.F1	SNORA23.q.R1	GATCTTGCTATCCACACAAACATCATGC	TCCAGAGACAACACTAGACCAGTACT
SNORA28	SNORA28.q.F1	SNORA28.q.R1	GGCAGATGATCAAAACTGTCTGACAC	AGTCTATATAACGGCTTGTCTCATGGG
SNORA34	SNORA34.q.F1	SNORA34.q.R1	AGACCAGCAGTTGTACTGTGGC	GCCATTCCCTACTGAGGTCCCA
SNORA44	SNORA44.q.F1	SNORA44.q.R1	GGGCTGTGGCTGGTCATAGC	AAAGCTGAGTGGCAGCTTGCAG
SNORA46	SNORA46.q.F1	SNORA46.q.R1	TCCCCATTCTTGGTTACGCTGT	TGTTCCTTAACTCTATACAGCAACAGCA
SNORA63	SNORA63.q.F1	SNORA63.q.R1	TAAGTGCTGTGTTGTCGTTCCCC	TATGAGACCAAGCGTCCCTGGC
SNORA64	SNORA64.q.F1	SNORA64.q.R1	AGTTGCACTTGGCTTCACCCG	GCACCCCTCAAGGAAAGAGAGG
SNORA68	SNORA68.q.F1	SNORA68.q.R1	GAATCACTGTTTCTTATAGCGGTGGTT	AAATTCACTTTGAGGGGCACGG
SNORD124	SNORD124.q.F1	SNORD124.q.R1	GGATGATGTTCCAGTTGAGACTCAAGAA	GGTCAGGGACCAAGTGGCTCC
SNORD13	SNORD13.q.F1	SNORD13.q.R1	AGCGTGATGATTGGGTGTTCATACG	CAGACGGGTAATGTGCCCACG
SNORD16	SNORD16.q.F1	SNORD16.q.R1	AATTTGCGTCTTACTCTGTTCTCAGC	TCAGTAAGAATTTTCGTCAACCTTCTGTAC
SNORD32A	SNORD32A.q.F1	SNORD32A.q.R1	AACATTCACCATCTTTCGTTTGAGTCTCAC	GTCTCAGAGCGGTGCATGGG
SNORD46	SNORD46.q.F1	SNORD46.q.R1	AAAAGAATCCTTAGGCGTGGTTGTG	CAGTCAGTGTAACTATGACAAGTCCTTG
SNORD67	SNORD67.q.F1	SNORD67.q.R1	GTTGCACACTGGTGGAGCCATG	GAGTCAGATGGCCCCTGTGC
SNORD83A	SNORD83A.q.F1	SNORD83A.q.R1	AGGCTCAGAGTGAGCGCTGG	GTTCTCAGAAGGAAGGCAGTAGAGAA
SNORD88C	SNORD88C.q.F1	SNORD88C.q.R1	AGCACTGGGCTCTGATCACCC	CCTCAGACCCCCAGGTGTCAA
SNORD89	SNORD89.q.F1	SNORD89.q.R1	ACAAGAAAAGGCCGAATTGCAGT	GAGGTCAGACTAGTGGTTCGCTT
VTRNA1-1	VTRNA1-1.q.F1	VTRNA1-1.q.R1	TCAGCGGTTACTTCGACAGTTCT	AGGACTGGAGAGCGCCCG

Table 2: Summary of primers used in qPCR analysis for quantification of ncRNAs transcripts' abundance. 25 snoRNAs, misc_RNAs and scaRNAs were chosen for qPCR analysis. Primer sequences were mapped against the human genome (hg38 version 85) to ensure that no complementarity could be found in other transcripts than the target's sequence.

2.2 Analysis of genomic data

Data obtained through sequencing methods has to be collected, and analyzed (Figure 5). The collection and further treatment is designed to remove and identify potential biases in the experimental design. Collection of the data was done through the set up of a server dedicated to data storage. The transfer was done through SSH protocol. Once the transfer was completed, a checksum verified that the data retained its integrity. The data was transferred to a dedicated Mammouth node, a supercomputer part of Compute Canada housed at the University of Sherbrooke, so that further processing of data would benefit from the advantages of a larger CPU memory allocation and parallel programming friendly environment.

The first step in a computational analysis of genomic sequencing data is to ensure that the sequencing procedure was successful and that the quality of the data is satisfactory. This step of Quality Assessment intakes the raw fastq files that are acquired from the sequencing facilities.

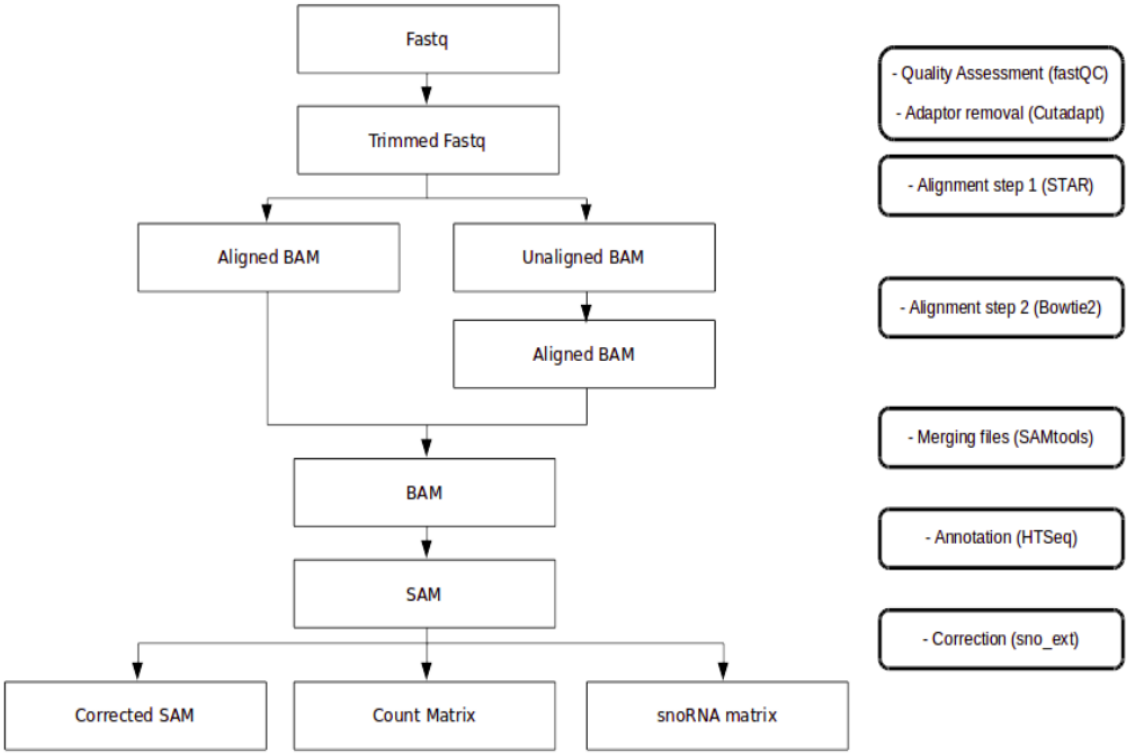


Figure 5: Overview of a pipeline for the analysis of genomic sequencing data. The data is treated to verify its quality and remove unwanted components (low quality reads and adaptors). The reads are then aligned and annotated to yield various matrices.

2.2.1 Quality Assessment

The fastq files were run through fastQC (version 0.11.5). The files were inspected for over representation of bases, lower quality bases, and abnormalities in the size of the reads. All files were judged to be of satisfactory quality.

2.2.2 Quality Treatment

Often the small RNAs species were smaller than the read length. As such, part or the entirety of the adaptor sequence would be present within the read. To palliate to the possibility of having an adaptor's sequence within the 5' or 3' of the read, the second step of the bioinformatics pipeline was to remove the adaptor sequences from within the reads.

The second step was also accompanied with the removal of much lower quality bases from the extremities of the reads. This was done through Cutadapt (version 1.19). Cutadapt is a stand-alone program primarily written in the python programming language with the alignment algorithm being implemented in C. It uses a mismatch threshold and does not require the user to type his own script and functions. As such, Cutadapt is a highly customizable and user-friendly program (Martin, 2011).

The adaptor sequences were given to Cutadapt:
GATCGTCGGACTGTAGAACTCTGAACGTGTAGATCTCGGTGGTCGCCGTATCATT
,
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGT
CTTCTGCTTG, TGGAATTCTCGGGTGCCAAGG and
GATCGTCGGACTGTAGAACTCTGAAC

Cutadapt's parameters were set for the read length threshold to be above 12, with base qualities (phred scores) being above 20, disallowing for indels (insertions or deletions), and allowing for wildcard (unidentified bases) characters. The other parameters remained to their default values which included the error tolerance at 10% of the adaptors' lengths. The overall command issued to Cutadapt was: cutadapt -a {adaptor_FWD} -A {adaptor_REV} --minimum-length 13 --no-indels --match-read-wildcards -q 20 -o {input_reads1}.fastq -p {input_reads2}.fastq {output_reads1}.fastq {output_reads2}.fastq >> {cutadapt_stats}.txt

2.2.3 Alignment

To identify known transcripts from a sequencing dataset, it is, first, common to select an appropriate genome and corresponding annotation. The cell lines being of the human ovarian cancer (SKOV3ip1), the human genome (hg38 version 85), comprising all canonical chromosomes but the Y, was extracted from Ensembl in the fasta format. As for the transcripts and gene annotations, we found out early that various agencies had annotated the human genome which allowed for a certain margin of variation in their respective annotation of transcripts and genes. As such, annotations of the human genome (gtf format) were pooled and repetitions were erased to create a more complete annotation file. The main human genome annotation, hg38 version 85, was extracted from Ensembl (Yates et al., 2016). The Ensembl annotation was supplemented with the genomic transfer RNAs, totalling 628, from the UCSC's GtRNAdb (Chan & Lowe, 2009). The Ensembl annotation was further modified by adding the snoRNAs, totalling 20, that were found missing from its annotations when crosschecking with RefSeq version 75 (O'Leary et al., 2016). In the goal of identifying variations in the global representation of various small ncRNAs, it was of the utmost importance to find a reliable mechanism by which we could identify ncRNAs. Many algorithms have been implemented in the past and they all achieved various levels of efficacy. However, they have their individual drawbacks such as under representation of smaller species, high run times and high memory allocation requirements. To palliate to these problematic circumstances, multiple alignment programs were used to achieve the most representative species detection possible.

In a first round, the alignment program STAR (version 2.5.1b) was used to capture most of the species. STAR is a stand-alone program written in C++. It used a precompiled index of the target genome which required the input of a gtf file containing the full genomic annotation prior to alignment (Dobin et al., 2013). The aforementioned gtf was fed into STAR (--runMode genomeGenerate, --runThreadN 44, --genomeDir {modified_hg38}.gtf, --genomeFastaFiles {genome}.fasta, --sjdbGTFfile {genome}.gtf, --sjdbOverhang 124) to build a large index (ensembl_star_index}.

The index was later used by STAR to produce the binary alignment map (BAM) files, a binary version of the sam formatted files. The STAR alignment was performed with the following parameters: --runMode alignReads --genomeDir {ensembl_star_index} --readFilesIn {output_reads1}.fastq {output_reads2}.fastq --runThreadN 45 --outReadsUnmapped Fastx --outFilterType BySJout --outStd Log --outSAMunmapped None --outSAMtype BAM SortedByCoordinate --limitGenomeGenerateRAM 250000000000 --limitIObufferSize 4000000000.

Once the first alignment completed, the BAM files contained a few thousand identifiable sequences that had not been mapped properly to snoRNAs. To adjust the BAM files, we used bowtie2 (version 2.2.4) which has the highest sensitivity to smaller (<50nt) species (Langmead & Salzberg, 2012). The first step was to generate an index file for bowtie2 using bowtie2-build and the fasta files containing the chromosomal sequences (Yates et al., 2016). The BAM files containing the unmapped reads were sorted and changed back to fastq format using bam2fastx. The resulting sorted fastq file was fed into bowtie2 with the parameters for local alignment of a minimal length of 13 bps using 48 processes and the human gtf file previously mentioned (bowtie2 --local -p 48 -q -x {bowtie2_index} -1 {unmapped1}.fastq -2 {unmapped2}.fastq -I 13 -S {htseq_annotated}.sam). Bowtie2 outputted a SAM file which was merged with the output from the mapping step with STAR using SAMtools (version 1.3).

2.2.4 Read Annotation

Once the second alignment step was completed and the reads were mapped to genomic intervals, these intervals were associated to genes and transcripts to estimate gene counts. This estimation can be obtained from a number of ways. One of the most common is to add a field in the mapping file indicating the gene / transcript found at the specified genomic interval. This process can become more involved when specific splicing variants of exons have to be measured. However, measuring exons' expression falls outside of this experiment's scope of interest. The program used in the annotation process had to be simple to use, to modify and it had to be well documented.

As previously mentioned, the most widely used programs handling reads annotation and counting were RSEM, HTSeq and bedtools (Anders et al., 2015; Li & Dewey, 2011; Quinlan & Hall, 2010). HTSeq (version 0.6.1p1) was selected because of the readability of its code, its wide use, its ability to handle paired-end data and the availability of its source code. HTSeq, itself, is a platform to execute common operations on sequencing data formats. The platform and its associated scripts come as a python package. The core principles of its HTSeq-count script were simple and could be modified by adding unto it.

HTSeq was fed the gtf annotation files previously mentioned, the uncompressed (and sorted by name) BAM files, a minimal quality score threshold and a mode to handle overlapping features (such as genes, etc). Through trial and error, it was determined that the default quality score for the alignment was too stringent if the reads mapping to ncRNAs with intergenic extensions or shifts were to be retained. As such, to allow extra sequences in nucleotides, the threshold was lowered to 5. The exact threshold limit was determined by progressively lowering the associated value and counting the total number of snoRNAs species detected in comparison to the bedgraph representation. Three modes are available to HTseq's analysis: union (default option), intersection strict, intersection non-empty. The goal of the HTseq is to discard reads that do not fall in known gene annotations while giving a basic annotation that will be modified further down the line. This relaxed discrimination process requires the broadest criteria to return the most annotated reads. As such, the intersection strict mode cannot be used as it discards every read exceeding annotations boundaries. The default mode was also rejected as it put multiple overlaps between annotations as ambiguous without accounting for the read's coverage. The command was given as such: htseq-count -f sam -r name -a 5 -m intersection-nonempty -o {outputted_annotated}.sam {htseq_annotated}.sam {modified_hg38}.gtf > {count_matrix_genes}.txt. The outputted files were an annotated sequence alignment map (SAM) file and a matrix describing the genes' accumulations.

2.2.5 Annotation Correction

The read annotation was incomplete as a high proportion of reads would be labelled as ambiguous or otherwise mislabelled. However, upon inspection, the largest part of the ambiguous reads were found to match snoRNAs (or other small ncRNAs) with annotations overlapping other RNA species' annotations. To bypass this issue, a python package was designed, sno_ext. It uses pandas matrices and parallel programming to speed up the annotation modification procedure. sno_ext was given the SAM files generated previously by the read annotation step and the modified human genomic annotation file previously mentioned. The exact command sent was: python3 ./pipeline.py {outputted_annotated}.bam {modified_hg38}.gtf snoRNA.

The corrective steps taken by sno_ext relied largely on a few heuristic principles. These assumptions were:

-Valid reads are aligned to annotated transcripts of known genes.

-Annotations with the longest overlap to reads were the origin of the aforementioned reads.

-Valid paired mates provide the exact genomic positions of both 5' and 3' ends (Figure 6).

The program used the provided formatted annotation to construct a list of possible annotations by extracting every position assigned to a gene and adding custom entries. These entries were created by combining neighbouring exons together in a process coined as "exon bridging". This bridging would discriminate between a read mapping to both ends of an exon-exon junction or one mapping to an element found within an intron. The bridging process was shortened by only including exons bordering ncRNAs. Once all the annotations computed, the files containing the annotated reads is processed. The reads are divided based on the previous steps in their treatment. HTSeq annotated files have a specific flag, a string of characters describing the read, stating how the file was processed and the associated gene annotation. This information allows the entries to be divided and treated exclusively when required as to not needlessly extend its runtime.

In this way, the reads identified as ambiguous are sequestered. The reads are given all the entries of overlapping annotations corresponding to their given HTSeq gene annotation. Each entry is scored based on the overlap between read and annotation coordinates (Figure 6). As an example, a read having a 3 nt difference on its 3' end with its 75 nt reference would net a score of 3/75 or 0.04. The lowest score is isolated and the read's tag is modified to fit the gene's entry. The resulting data structure contains all reads to be further classified and treated. First, the reads are grouped and counted. The counts and genes are presented as a matrix containing additional information such as individual CPM, TPM, gene's type and length. The matrix is further collapsed based on the gene's type. In this instance, the counts and gene types are outputted as a matrix describing the abundance (CPM and TPM) of each type. A supplementary matrix is generated for snoRNAs. This table contains each snoRNAs name with each unique detected isoform coordinates and sequence. The last script could easily be adapted for any ncRNAs reads shorter than the annotation's length.

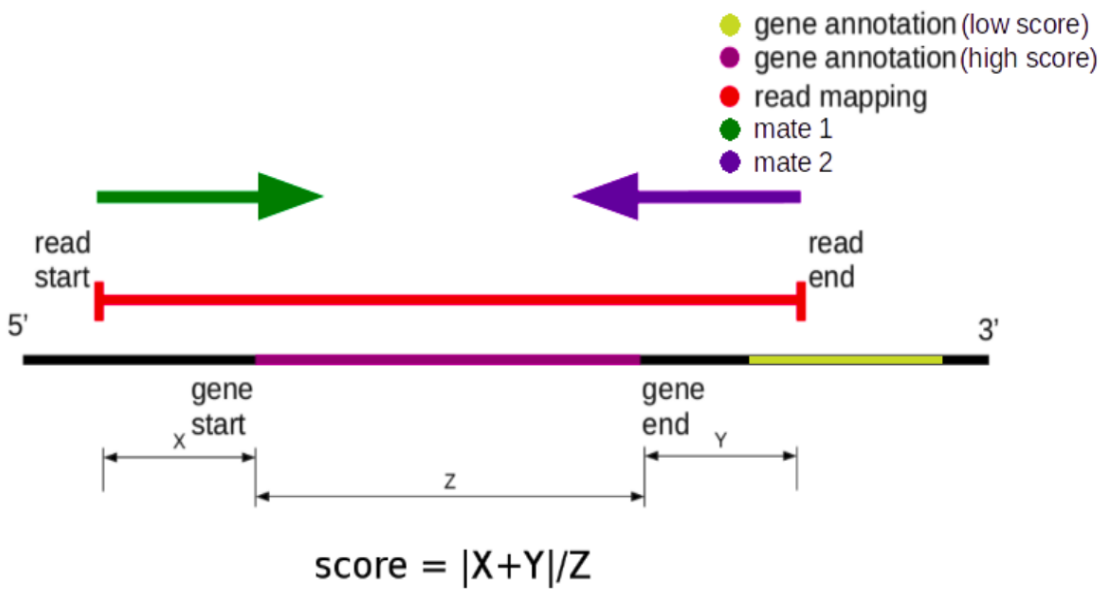


Figure 6: Scoring schema for gene identification. The scores are calculated by taking the areas of non overlap (X & Y) between gene and read mapping and normalizing by the gene annotation length (Z) between all possible transcripts. The length of the area of non overlap is the sum of the difference between mates' ends and annotation ends.

The key to sno_ext's processing power is the use of vectorization through pandas. Pandas were developed to perform operations through matrices' rows and columns rather than treating individual elements. By combining pandas and splitting tasks between children processes, we achieve a shorter runtime. However, the runtime remains quite high, approximately 2 days for the bigger datasets (~50GB), as constructing the annotation requires preprocessing.

3. RESULTS

3.1 Quality assessment

Once the sequencing of the submitted libraries was completed, the datasets were stored on a remote server specifically allocated to house the large (>100 GB) quantity of data generated. The data was transferred safely through sftp transfer to our local machines. The integrity of the data had to be verify. As such, checksum were performed on both the data before and after the transfer. Once, the sets were validated as unaltered by the transfer, an initial analysis of the quality of the sets was performed by FastQC. The FastQC analysis incorporated evaluations of the sequence quality, GC content, length distribution, and representation of sequences. From this selection, the sequence and base quality scores were of interest (Figure 7). The overall median quality of the reads for the VUSs datasets is maintained above 30 (base call accuracy of 99.9%) until 90 nt from the beginning of the read whereas the TGIRT-based sequencing datasets dip below a phred score of 30 between 120-130 nt from the start. These constantly high quality values validate the continuation of the analysis.

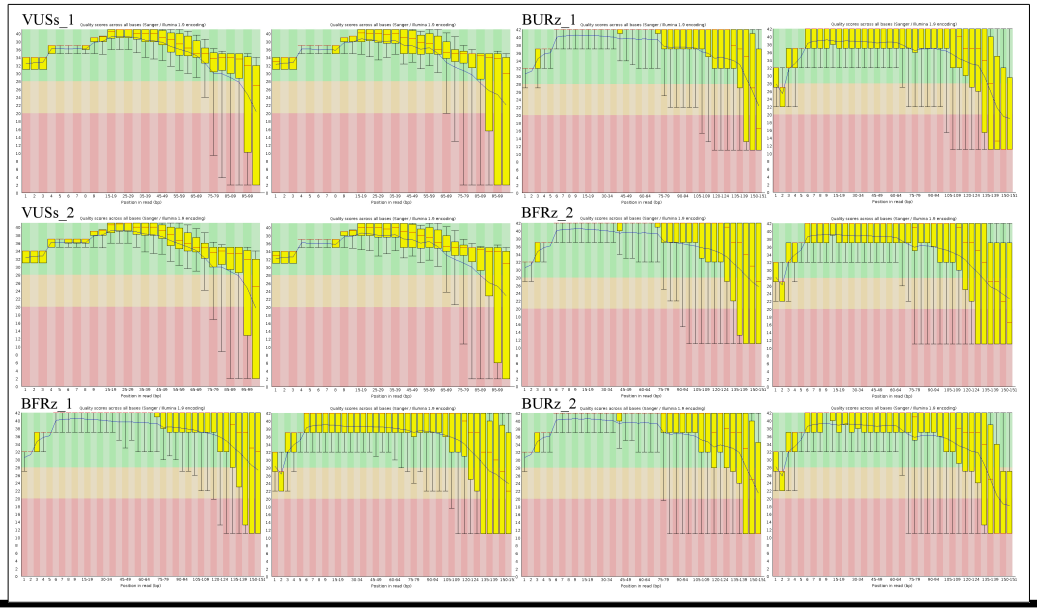


Figure 7: Global assessment of per base quality (phred score) in studied sequencing datasets before processing. A general overview of the datasets using FastQC shows that, for B*Rz sets, the average quality dips below 30 between 120-130 nt while it drops below the same threshold between 90-100 nt for VUSs sets. For each dataset, a pair of plots are displayed, the leftmost corresponds to the forward read while the rightmost, is the reverse read of the pair.

3.2 Adaptor removal efficacy

	VUSs_1	VUSs_2	BFRz_1	BFRz_2	BURz_1	BURz_2
Raw reads	18,530,306	16,429,902	63,121,910	45,630,572	53,160,922	41,913,968
Trimmed adaptor 1	14,104,508 (76.1%)	12,695,697 (77.3%)	45,429,616 (72.0%)	38,602,961 (84.6%)	43,803,235 (82.4%)	37,569,834 (89.6%)
Trimmed adaptor 2	13,574,485 (73.3%)	12,224,598 (74.4%)	53,031,229 (84.0%)	39,642,061 (86.9%)	48,215,581 (90.7%)	38,716,344 (92.4%)
Removed short sequences	1,409,395 (7.6%)	1,074,613 (6.5%)	2,143,156 (3.4%)	4,157,736 (9.1%)	3,678,978 (6.9%)	4,482,941 (10.7%)
Adaptor free reads	17,120,911 (92.4%)	15,355,289 (93.5%)	60,978,754 (96.6%)	41,472,836 (90.9%)	49,481,944 (93.1%)	37,431,027 (89.3%)

Table 3: Cutadapt summary of processed reads. Datasets have between 16 and 64 millions of reads before treatment. Adaptors are detected in most of the reads (72-93%). The datasets retain most of their reads with only 3-11% removed because of their length. All percentages are given relative to the reads count before treatment.

Once the quality of the reads was ascertained, the adaptors were removed from the reads through cutdadapt (Table 3). Cutadapt identified and excluded the adaptors in a majority of the reads. In the size selection-based sets (VUSs), 77% of the reads had adaptors removed. The BURz and BFRz sets had comparable rates of adaptor exclusion ranging from 72% to ~90%. The removal of very small reads rejected between 1 and 4.5 millions reads for being under the 13 nt threshold. However, these numbers make up a minority of the total reads which remained over 90% of their initial count. Once purged of non-genomic components, the reads can be mapped to the human genome with the expectation of high mapping scores. To this end, alignment programs were compared to identify the optimal configuration of the aforementioned software.

3.3 Aligner performance assessment

	STAR	Bowtie2	Tophat2	Kallisto
Speed ranking	2	3	4	1
Memory usage ranking	4	2	3	1
Accuracy ranking	2	1	4	3
Splice awareness	Yes	No	Yes	No

Table 4: Performance ranking of widely used mapping programs based on literature. Ranking is assessed based on past benchmarking experiments, listing programs from best (1) to worst (4). Kallisto is the faster and less memory consuming program. STAR is more accurate, especially in detecting small species. STAR is also splice aware allowing for the identification of transcripts. Tophat2 is slower than the other aligners. Bowtie2 requires less memory than others, but lacks splice awareness.

By examining the literature and benchmarking articles published, it is possible to rank the various aligners previously described (Table 4). Kallisto was estimated to be many fold faster, often taking less than 10 minutes where other take hours. The second faster program is STAR which was estimated to be 4 times faster than Tophat2 on test sets. Bowtie2 was also faster than Tophat2, sometimes reaching half its compute time. Kallisto was also estimated to require the less memory as the index contains non repeated transcriptomic sequences as opposed to the more sizable genome. The second less RAM consuming program is bowtie2 with its human genome index taking ~3GB while STAR is said to require at least 30 GB. Tophat2 is a modification to bowtie2 to make it splice aware, as such, it scales in memory usage similarly to bowtie2. Accuracy evaluated on simulated sets composed of mRNAs indicated that bowtie2 was the most accurate to evaluation abundance. STAR and Kallisto were close second and third, respectively. Tophat2 came in last with the highest variation in median estimation of abundance. Examining the strengths and weaknesses of individual programs, STAR was used conjointly with bowtie2 to perform the mapping. Once a pertinent workflow selected, all datasets were mapped to the human genome, hg38 version 85.

3.4 Cumulative mapping report

	VUSs_1	VUSs_2	BFRz_1	BFRz_2	BURz_1	BURz_2
Adaptor free reads	17,120,911	15,355,289	60,978,754	41,472,836	49,481,944	37,431,027
STAR aligned reads	13,672,760 (79.9%)	12,227,416 (79.6%)	53,472,962 (87.7%)	37,326,743 (90.0%)	43,141,208 (87.2%)	33,850,447 (90.4%)
Bowtie2 saved reads	1,015,612 (5.9%)	704,807 (4.6%)	5,390,631 (8.8%)	3,092,595 (7.5%)	5,401,022 (10.9%)	3,065,350 (8.2%)
Total aligned reads	14,688,372 (85.8%)	12,932,223 (84.2%)	58,863,593 (96.5%)	40,419,338 (97.5%)	48,542,230 (98.1%)	36,915,797 (98.6%)

Table 5: Adaptor free reads mapping summary to human genome (hg38 version 85) by STAR and bowtie2. The size selection datasets show a lower read mapping average (85.0%) compared to the more recent datasets generated using TGIRT (~98%). STAR maps over ~80% of the reads, while bowtie2 rescues anywhere between 4 and 11%. All percentages are given as fractions from their associated number of adaptor-free reads.

Once cleared of the non genomic adaptor sequences, the reads were mapped to the human genome (Table 5). This step was divided into two successive alignment phases. The first, conducted by STAR, identified a majority of reads as concordant and mapping to the genome. In the VUSs datasets, ~80% of reads were mapped within the first mapping round while the BFRz and BURz sets were, on average, mapped conclusively in ~90% of the reads. The following phase, using bowtie2 to map the unaligned reads, rescued anywhere from 4 to 11% of the total reads. The overall mapping, in the VUSs sets, was lower than the others by more than 10%. However, the overall alignment in each of the set remains higher than 84%. Once mapping was completed, the reads had to be associated with their respective genes through the process of annotation. First, however, a few programs to annotate reads had to be compared to identify the optimal available method.

	HTSeq	RSEM	bedtools
Speed ranking	1	3	2
Memory usage ranking	2	1	3
Ease of modification ranking	1	3	2

Table 6: Performance ranking summary of widely used annotation programs based on literature. Ranking is assessed based on past benchmarking experiments, listing programs from best (1) to worst (4). RSEM is the fastest and lower memory requirement annotation method. HTSeq is the second while bedtools is the slower and bulkier of the three. The ease of modification comes from language and structure of the code. RSEM would be the most difficult to modify, while HTSeq would be the easiest.

The programs available to annotate the mapped reads vary in their speed of execution, the CPU requirements and the ease with which someone might be able to modify them to fit better to specific analysis methods (Table 6). HTSeq was the fastest of all tested programs, able to return gene counts in mere minutes. RSEM and bedtools were pretty closely matched, both requiring more than an hour to run to completion. However, bedtools was a bit faster in treating the data. The programs requirements in term of memory allocation somewhat fluctuated. RSEM appeared to fare better, showing half the memory consumption that either bedtools or HTSeq exhibited.

The ease of modification is a subjective ranking based on an assessment of how difficult it would be to change features found within the source code. HTSeq is written in python and the code is easily readable making it the easiest code to alter. Bedtools multicov's body is written in C++, making readable, but long to change conclusively as the classes are scattered throughout the Bedtools files. RSEM is ranked third because its written in C++, but also contains a threshold below which normalization encounters issues. Taking into account the previously mentioned criteria, HTSeq was selected to annotate reads. Once the annotation method selected, the data was run through the pipeline.

3.6 Total annotation report

	VUSs_1	VUSs_2	BFRz_1	BFRz_2	BURz_1	BURz_2
Total aligned reads	14,688,372	12,932,223	58,863,593	40,419,338	48,542,230	36,915,797
Multimapped reads	3,788,392 (25.8%)	3,066,230 (23.7%)	8,899,671 (15.1%)	6,887,064 (17.0%)	6,295,219 (13.0%)	5,233,508 (14.2%)
Ambiguous reads	4,404,581 (30.0%)	3,965,356 (30.7%)	5,213,830 (8.8%)	3,465,202 (8.6%)	5,416,010 (11.2%)	4,311,185 (11.7%)
No feature reads	1,833,017 (12.5%)	1,613,129 (12.5%)	5,102,860 (8.7%)	3,989,878 (9.9%)	6,001,329 (12.4%)	4,582,969 (12.4%)
Annotated reads	4,662,382 (31.7%)	4,287,508 (33.2%)	39,647,232 (67.4%)	26,077,194 (64.5%)	30,829,672 (63.5%)	22,788,135 (61.7%)
sno_ext rescued reads	4,404,581	3,965,356	5,213,830	3,465,202	5,416,010	4,311,185
Total annotated reads	9,324,764 (63.4%)	8,252,864 (63.8%)	44,861,062 (76.2%)	29,542,396 (73.1%)	36,245,682 (74.7%)	27,099,320 (73.4%)

Table 7: Read annotation summary by HTSeq and sno_ext. Most reads are annotated successfully (>63%) with datasets generated from TGIRT having a higher percentage of annotated reads (~74%) compared to their size selection counterparts (~63%). Using sno_ext rescued all ambiguous reads accounting for ~ 4 millions reads in each dataset, by assigning gene identities. All percentages are fractions of their total aligned reads.

The annotation process is composed of multiple steps, the annotation through HTSeq and the correction to the annotation from sno_ext (Table 7). The HTSeq annotation could not rescue reads that were mapped to multiple sites within the target genome. These multimapped reads were excluded from the annotation and represented a significative part of the total reads number. Multimapped reads accounted for 24% of the total depth of the VUSs sets whereas the other sets generated through the TGIRT protocol had between 13 and 17%. Reads that were mapped to genomic position that did not have features were also excluded from further processing. Only 8 to 13% of reads were identified as having no feature. A final category, the ambiguous reads, were discarded by HTSeq. These reads fell within a genomic interval shared by two or more features. A large portion of reads fell in that category, from 9 to 31%. The higher proportion of ambiguous reads was found in the VUSs sets.

The overall annotation with the minimal version of HTSeq identified between 32 and 68% of the total number of reads. However, using sno_ext rescued all ambiguous reads which brought the overall annotation between 64 and 76% of the starting read number.

3.7 Overall reads distribution within RNA families

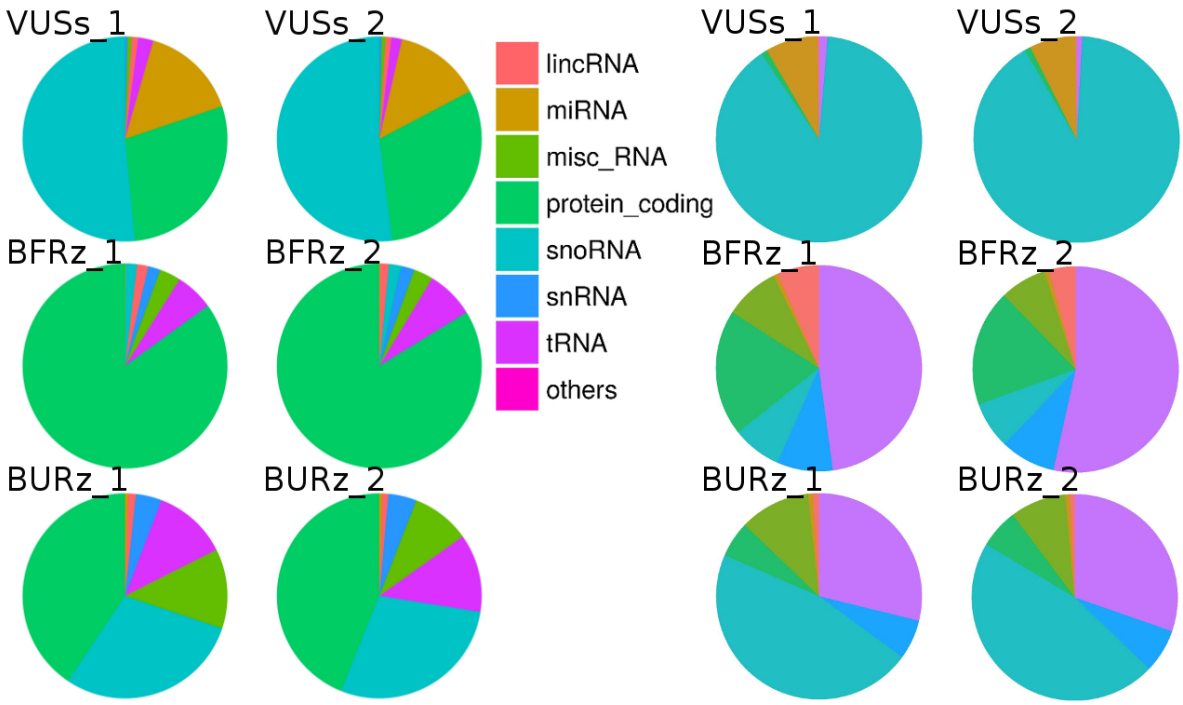


Figure 8: Global relative read expression (%) in CPM (left) and TPM (right) of the RNA families from the HTSeq analysis before corrections from sno_ext. Replicates show conserved patterns of expression. Size selection experiments are composed primarily of snoRNAs (CPM:~47%, TPM:~81%), followed by transcripts mapped to protein coding genes (~26%) but normalizing for transcripts' length reduces the latter's abundance to less than 1% (~0.8%) .miRNAs represent ~14% of reads in size selection experiments which when normalized for length becomes the second most abundant with ~7% of global species detection. In fragmented sets made with TGIRT, the most abundant (CPM) families are protein coding RNAs (~82%), tRNAs (~6%) and miscRNAs (~3%). However, when normalized for length (TPM), the most abundant species are tRNAs (~50%) and protein coding (~18%). In the unfragmented sets made by following the TGIRT protocol, the most abundant families (CPM) are protein coding RNAs (~35%), snoRNAs (~25%), tRNAs (~10%) and misc_RNAs (~10%). However, normalizing for transcripts' length, the most abundant species become tRNAs (~27%), snoRNAs (~43%), and miscRNAs (~9%).

Looking at the distribution of the reads annotated by the default HTSeq-count analysis (Figure 8) in the VUSs sets, the initial distribution of RNA families showed a large proportion of reads, 47%, mapping to snoRNAs. The other major detected families were protein coding and miRNAs with 26% and 14%, respectively. The BFRz sets had reads primarily mapping to protein coding transcripts (82%), tRNAs (6%) and miscellaneous RNAs (3%). The unfragmented BURz sets' annotation indicated that the most abundant RNA families were protein coding RNAs (35%), snoRNAs (25%), tRNAs and miscRNAs (both at 10%).

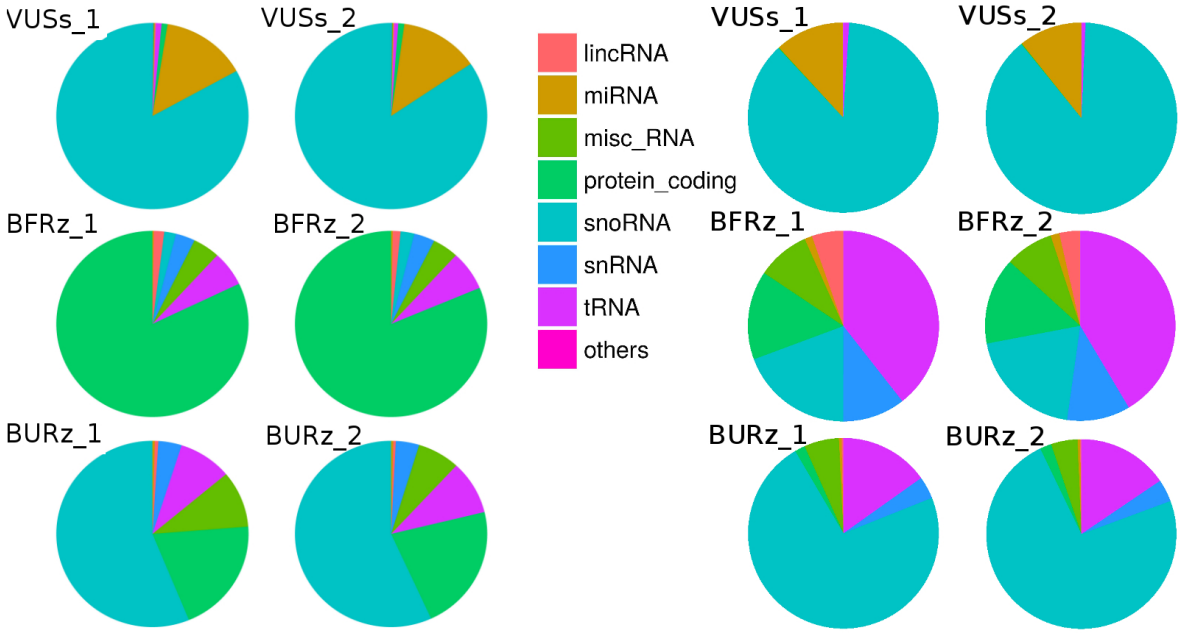


Figure 9: Global relative read expression (%) in CPM (left) and TPM (right) of the RNA families from the HTSeq analysis after corrections from sno_ext. Replicates show conserved patterns of expression. Size selection experiments are composed primarily of snoRNAs (CPM:~81%, TPM:~87%), followed by miRNAs (CPM:~14%, TPM:~13%). In fragmented sets made with TGIRT, the most abundant reads are mapping to protein coding RNAs (~78%), tRNAs (5.8%) and misc_RNAs (4.3%). When normalizing for species size (TPM), the smaller species are more represented (tRNAs: ~36%, snoRNAs: ~17%, snRNAs: ~10%, miscRNAs: ~8%, miRNAs: ~1%) than the bigger species (protein coding: ~14%, lincRNAs: ~4%). In fragmented sets, snoRNAs' abundance remained mostly unaffected (1.9%) when compared to VUSs. In the unfragmented sets made with TGIRT, the most reads mapped to snoRNAs (~52%), protein coding RNAs (~18%), tRNAs (~9%) and misc_RNAs (~8%). Normalizing for size, the most represented species are snoRNAs (~71%), tRNAs (~15%), miscRNAs (~5%) and snRNAs (~4%) while bigger molecules, protein coding species, are less represented (~2%).

On the other hand, after the correction from sno_ext to the annotation provided by HTSeq, the ratios between the various RNA families shifted (Figure 9). In the VUSs, the snoRNAs remained the most detected family, but its overall representation increased from 47% to 82%. The second most abundant family, the protein coding RNAs, after correction, made up less than 1% of the reads. The overall abundance of miRNAs remained at 14%. In the BFRz sets, the abundance for every family remained comparable whether the correction was added or not. The most abundant families did not vary more than 5% from before the sno_ext correction.

The most represented family were the protein coding RNAs (78%), with other families being found at low levels like tRNAs (6%) and miscellaneous RNAs (4%). As expected, the unfragmented sets, BURz, were highly affected by the correction from sno_ext because of wrongful gene assignment from HTSeq. The snoRNAs more than doubled in abundance, making them the most represented, with 52% of reads mapping to snoRNAs. The other most abundant families were the protein coding RNAs, which had reads reassigned to snoRNAs lowering their global representation to 18%, tRNAs and miscellaneous RNAs which remained in similar proportions after the correction (9% and 8%, respectively). SnoRNAs were further divided into family to inspect their individual abundance fluctuations and to examine H/ACA to C/D box snoRNAs ratios.

3.8 Overall reads distribution within snoRNA families

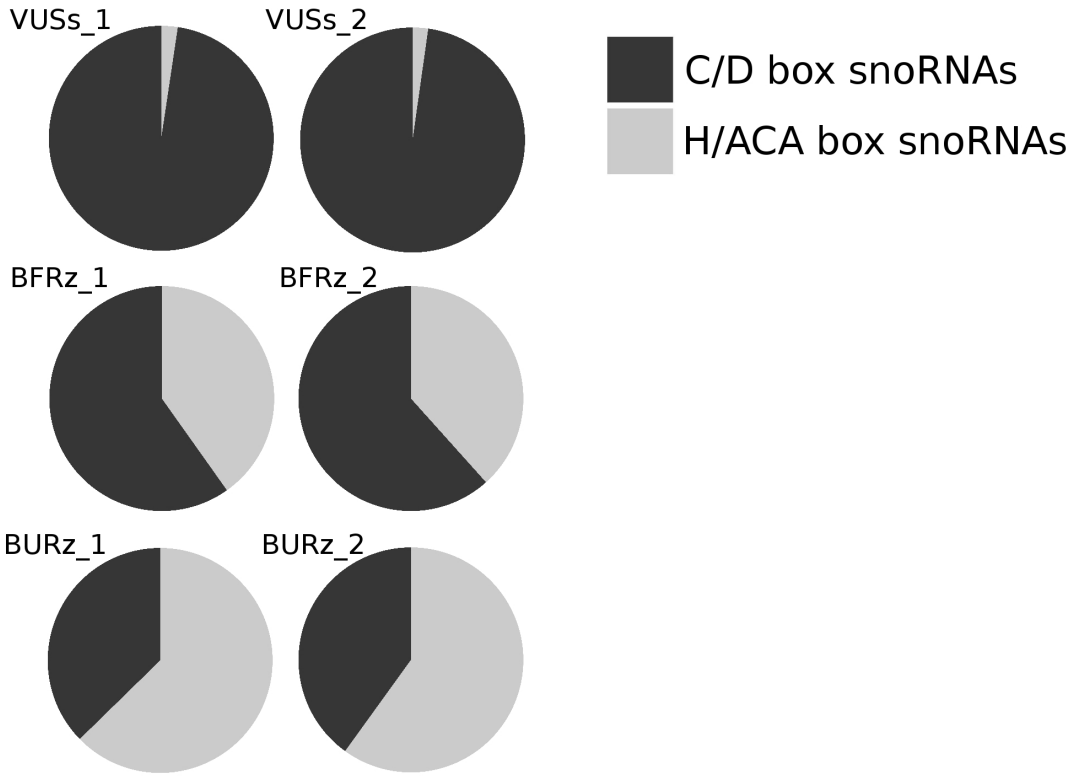


Figure 10: Relative expression (CPM) of the two snoRNAs families (H/ACA box & C/D box) in sequencing sets generated by size selection (VUSs) and the TGIRT method (B*Rz) before correction by sno_ext. The different sequencing replicates exhibit conserved expression patterns. The data produced by the size selection protocol had a higher proportion of reads mapping to C/D box snoRNAs (~98%), whereas the other sets' ratios were closer to being evenly distributed (60% and 37% for BFRz and BURz, respectively).

The snoRNAs were the most affected by the correction from sno_ext. Taking a closer look at the snoRNAs families (Figure 10), the C/D box snoRNAs were highly represented within the libraries generated through the standard size selection protocol (98%) as opposed to the ones generated through the TGIRT protocol (B*Rz) which had a higher proportion of H/ACA box snoRNAs (60% and 37% for BFRz and BURz, respectively). Such an overwhelming representation of C/D snoRNAs had been previously catalogued (Deschamps-Francoeur et al., 2014).

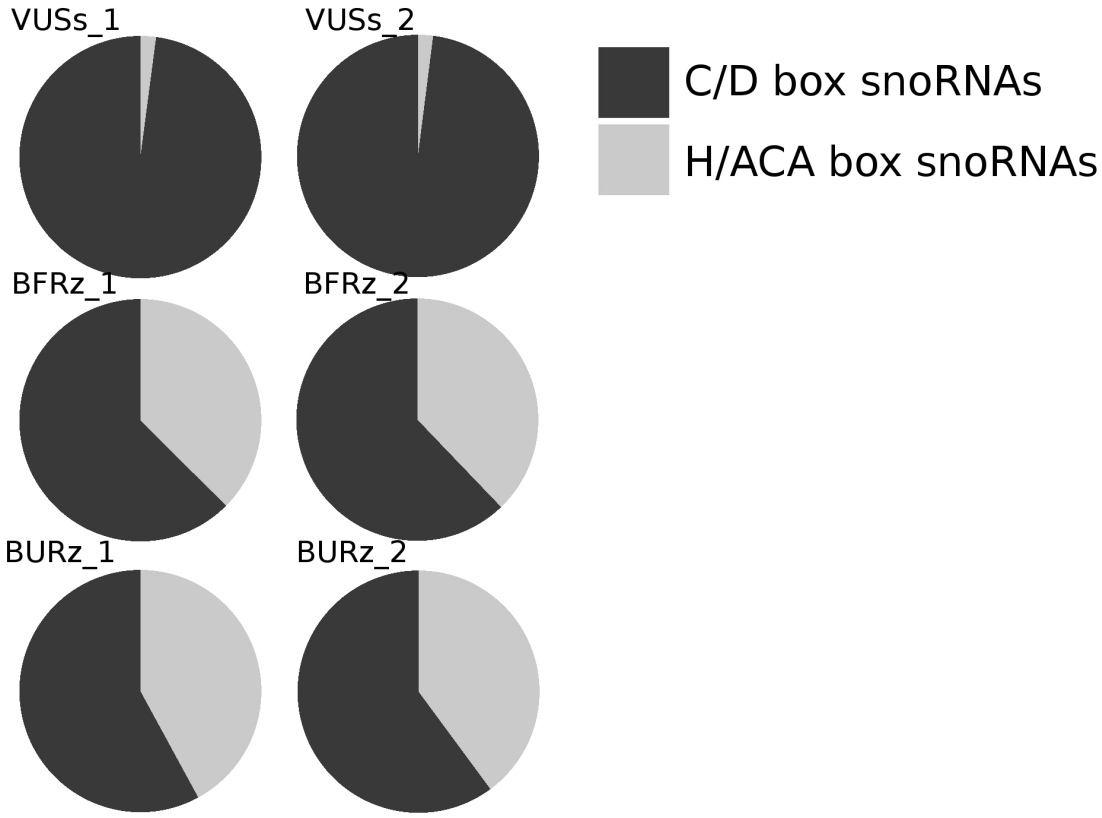
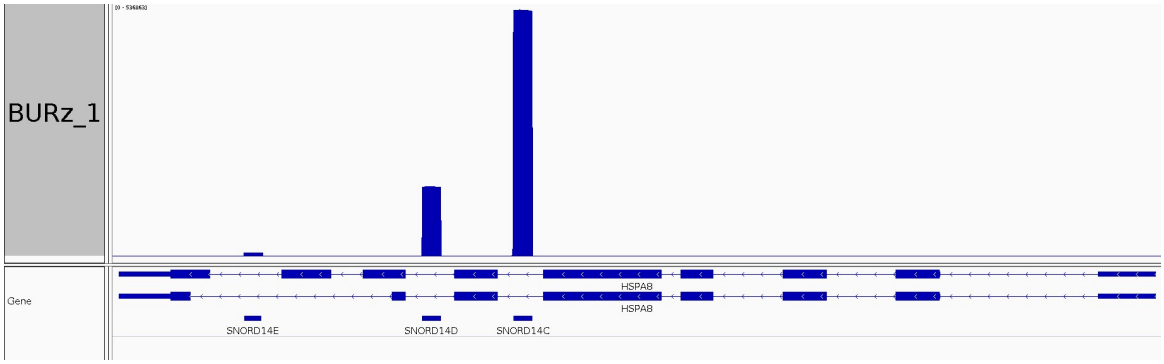


Figure 11: Relative expression (CPM) of the two snoRNAs families (H/ACA box & C/D box) in sequencing sets generated by size selection (VUSs) and the TGIRT method (B*Rz) after correction by sno_ext. The different sequencing replicates exhibit conserved expression patterns. The data produced by the size selection protocol had a higher proportion of reads mapping to C/D box snoRNAs (~98%), whereas the other sets' ratios were closer to being evenly distributed (63% and 58% for BFRz and BURz, respectively).

The correction from sno_ext, affected the overall representations of both snoRNAs families in a few sets (Figure 11). In the BURz datasets with the sno_ext modified HTSeq analysis, the C/D box snoRNA family was found to be more abundant than the standard HTSeq analysis would have led to believe, as it was increased to 58% from its original 37%. This trend was expected as BURz sets have the widest diversity of snoRNAs species. Thus, the other sets remained mostly unaffected by the sno_ext correction as snoRNAs species were not found to have a similar species' diversity. This correction affected heavily some of the datasets, individual accumulation profiles had to be examined to highlight scenarios where reads were rescued by sno_ext.

3.9 Effects of sno_ext's correction on data distribution



	HSPA8 count	SNORD14E count	SNORD14D count	SNORD14C count
HTSeq	539,855	7,936	0	0
HTSeq+sno_ext	1,637	7,944	151,811	536,871
RSEM	236,133	2,739	59,546	28,455
Raw count	2,254	7,882	151,262	536,865

Figure 12A: Species abundance and read mapping according to different annotation protocols in the BURz set for HSPA8 and snoRNAs found within its introns. The bedgraph indicated that SNORD14E, SNORD14D, SNORD14C had accumulations of 7882, 151262 and 536865 reads, respectively. HTSeq annotation had indicated more reads to the protein coding gene (539855) than its associated snoRNAs (7936). HTSeq accompanied with sno_ext returned more reads associated to snoRNAs (696626) than its host genes (1637). RSEM annotation had both a high of reads annotated to snoRNAs (90740) and the host gene (236133).

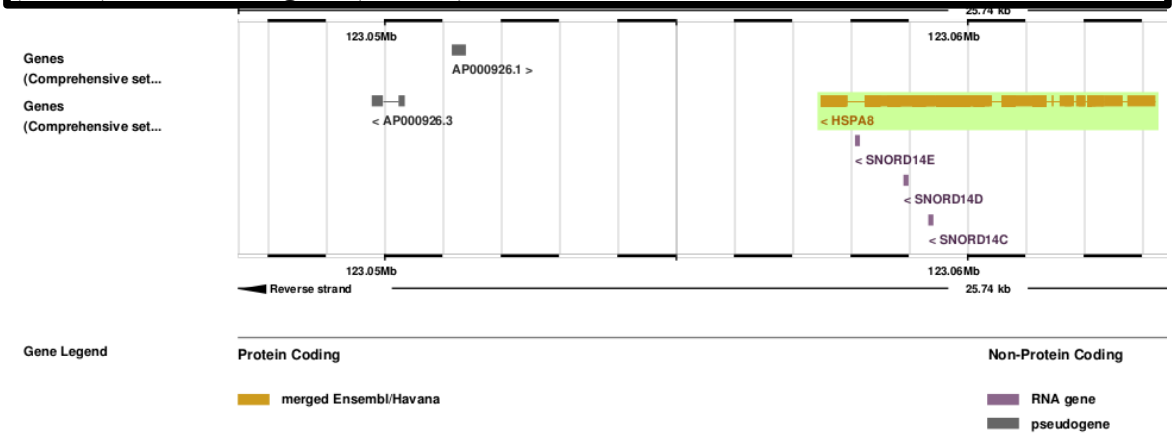


Figure 12B: Ensembl genome view of gene annotations mapping to chromosomal position of HSPA8 gene (Kb). Overlapping all gene annotations reveals multiple possible overlaps between alternative exons and ncRNAs. In this situation, HSPA8 annotations overlap SNORD14C and SNORD14D.

The correction from sno_ext redirected reads attributed to protein coding transcripts by the intersection-nonempty HTSeq to snoRNAs. The 539855 reads coming from the HTSeq analysis were reassigned to its snoRNAs (Figure 12A). The host gene, HSPA8, was found to have alternative exons that overlapped non-coding RNA annotations (Figure 12B). An alternative method for the annotation, RSEM, was shown to have snoRNAs counts fall below the values perceived from a visual assessment of the bedgraphs. This example is far from an isolated case as between 40 and 90% of the detected snoRNAs are affected by the correction from sno_ext. Since the data before and after correction was shown to have such large disparities, qPCR experiments with selected ncRNAs were carried out to identify the most adequate annotation method.

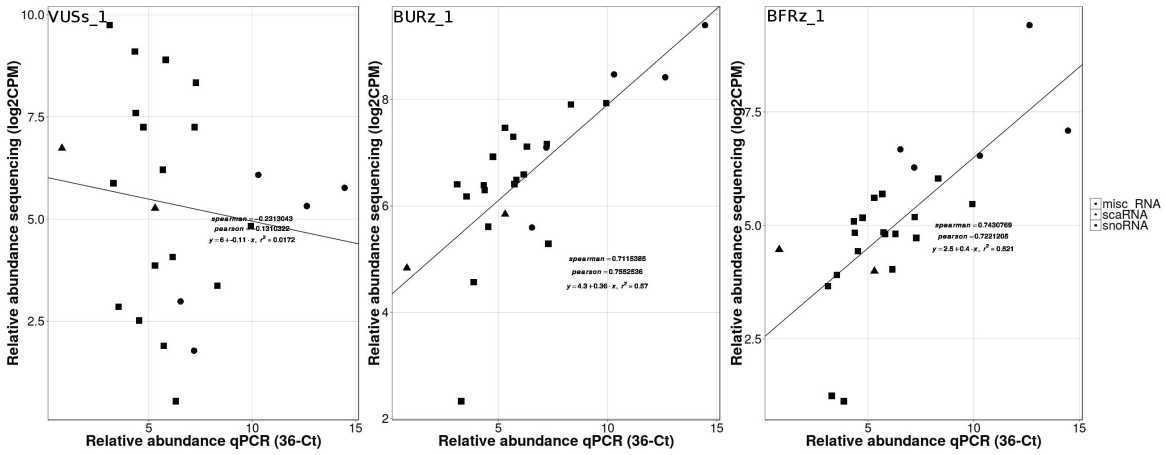


Figure 13: Correlation between quantification of ncRNAs from qPCR and sequencing in VUSs, BURz and BFRz sets after correction from sno_ext. The Spearman and Pearson correlations between the abundance values obtained for selected snoRNAs, scaRNAs and miscellaneous RNAs found in the VUSs set were -0.2313 and -0.1310, respectively. These correlations, for the same species, were increased in the BURz (0.7115 Spearman and 0.7553 Pearson) and BFRz (0.7431 Spearman and 0.7221 Pearson) datasets.

	HTSeq		HTSeq + sno_ext	
	Spearman coefficient	Pearson coefficient	Spearman coefficient	Pearson coefficient
VUS_1	-0.1200	-0.0250	-0.2313	-0.1310
VUS_2	-0.1308	-0.0335	-0.2417	-0.1441
BURz_1	0.4832	0.3590	0.7115	0.7553
BURz_2	0.4427	0.3463	0.6746	0.7201
BFRz_1	0.5775	0.4663	0.7431	0.7221
BFRz_2	0.5878	0.4745	0.7400	0.7024

Table 8: Correlation between quantification of ncRNAs from qPCR and sequencing in VUSs, BURz and BFRz sets. Pearson and Spearman correlations were calculated on all sets treated in this study. The VUS sets' coefficient are low, between 0.025 and 0.25, whether or not the sno_ext correction is used. The BURz and BFRz sets' coefficients are increased with the use of the correction from sno_ext.

Inspection of the species accumulation through sequencing alone could not help in ascertaining the validity of the accumulation profiles. To this end, a subset of ncRNAs were selected for the diversity of their predicted accumulation through sequencing. The selected species were tested through qPCR and correlated to their perceived count in the sequencing experiments. The values obtained from qPCR and sequencing had to be adapted to a linear model. These values were then correlated to one another to verify the presence of trends. In the VUSs sets, the qPCR correlated poorly with the sequencing data (Figure 13 and Table 8). The Spearman and Pearson coefficients for these sets remained between 0.10 and 0.25 (Table 8). On the other hand, the correlation coefficients (Spearman and Pearson) when comparing sequencing data from the BURz datasets to the aforementioned qPCR were between 0.70 and 0.75. The VUS sets' coefficients are multiple folds lower than any other coefficients. All coefficients are improved by the correction made by sno_ext. BURz and BFRz datasets' coefficients are increased between 0.15 and 0.37.

3.10 Addressing sequencing biases

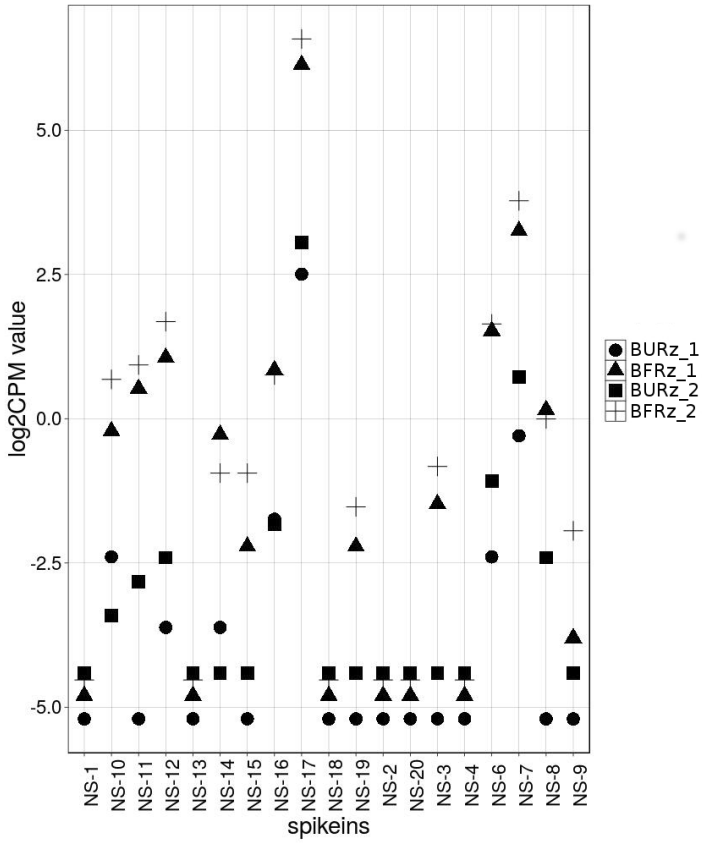


Figure 14: Assessment of spike-ins composition as a factor affecting abundance (log2CPM) in the datasets generated by the TGIRT protocol. Most spike-ins were not detected. The detection patterns were conserved between replicates.

To further address sequencing biases, the TGIRT sequencing datasets had been supplemented with ERCC spike-ins. Two sets of spike-ins were added to the sequencing libraries, the size-range quality control (SRQC) and the external reference for data normalization (ERDN). The spike-ins either varied in sequence composition (ERDN) or length (SRQC) (Locati et al., 2015). The ERDN spike-ins, all 19 oligoribonucleotides labelled NS, were detected at low levels (Figure 14). The most abundant was NS-17 (200-1000 reads) with an associated CPM from 6 to 95. The second most represented was NS-7 (50-300 reads) with an associated CPM from 6 to 8 times lower than NS-17. The wide fluctuations between spike-ins was not anticipated. These changes in abundance between spike-ins, although unexpected, only show that Lambowitz's protocol has room for further improvement in small RNA species representation.

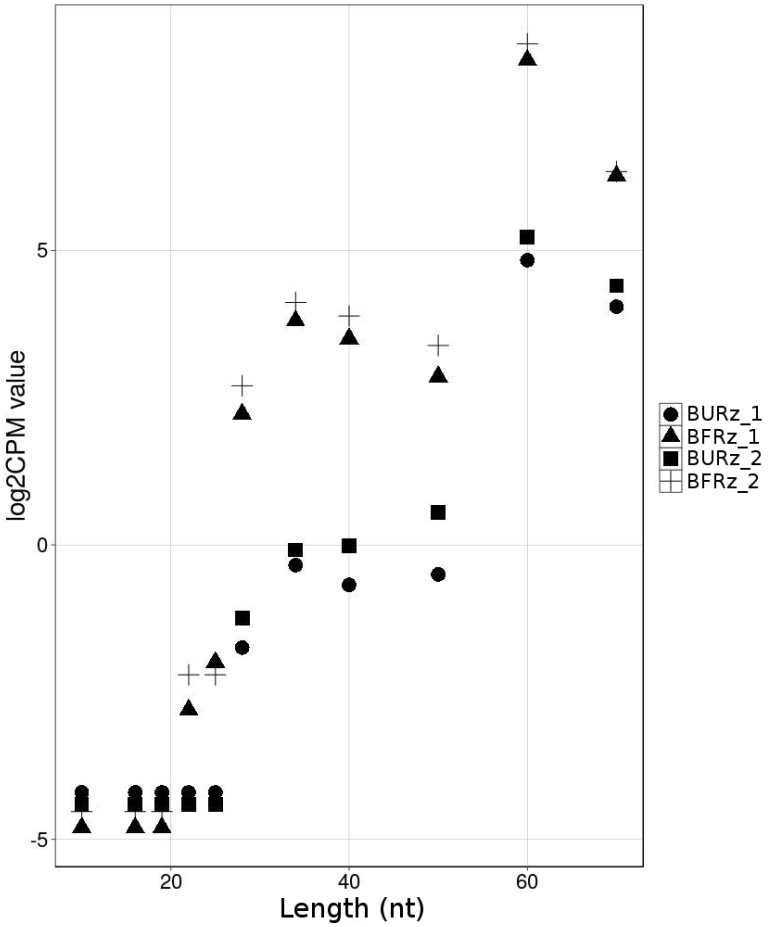


Figure 15: Spike-ins length (nt) correlated to their distribution (log2CPM) in all sets produced through the TGIRT protocol. Lower molecular weight (<20 nt) spike-ins are not detected. The spike-ins have linear upward trend where the most abundant species are 60 nt (29-364 CPM).

The second batch of spike-ins, SRQC, was made to assess the presence of any bias on transcript's length. The smaller spike-ins, shorter than 20 nt, were not detected (Figure 15). The spike-ins started to accumulate after 20 nt in a constant linear fashion reaching its apex at 60 nt where 795 to 8387 reads could be detected. The measures collected correlated quite well between replicates, however fluctuations of many folds are visible between fragmented and unfragmented samples. The overall positive correlation between abundance and molecule's length was not expected as it shows under-representation for low weight molecules.

4. DISCUSSION

4.1 Preparation of the data prior to analysis

A preliminary read inspection (Figure 7) revealed through their individual reads profiles, that the data is of high quality. A few have punctual drops in quality as shown in the plot for the reverse reads of BFRz_1, where, at the position of the second nucleotide, the median drops to 27. However, such events have been detected previously and do not affect the overall ability to continue with the analysis ("System User Guide," 2007). The global quality of the datasets is quite high, with sets created through the bacterial protocol retaining higher scores longer than their viral counterparts. The difference can be attributed to the advancements in sequencing in the past years which yield consistently more precise data. The long stretches of high quality bases found within the reads gives a full coverage of most snoRNAs detected.

However, the reads' coverage also displayed flanking sequences. These sequences could not be found within the human genome. These stretches were the adaptors sequences retained in the reads. The standard Illumina sequencing pipeline would remove these non-genomic parts, but the adaptors used within these experiments were not found within their listings. Thus, these bases had to be trimmed out before proceeding forward to positively identify the transcripts sequenced. Cutadapt identified and removed both adaptors in over 70% of reads (Table 3). A low proportion of reads, between 3 and 10%, were found to be shorter than 13 nt. These reads were removed since they could map to multiple sites in the genome which would impede drawing strong conclusions. We are left with over 90% of the initial reads count.

4.2 Mapping reads to the human genome

The remaining 90% of the reads had to be mapped to the human genome. The alignment process had to be optimized to fit the widest spectrum of molecules as possible. The literature was scoured to compare the most popular methods of genomic alignment. The widely used programs identified were kallisto, tophat2, bowtie2 and STAR.

Comparing the notes given through benchmarking experiments, the performances from each program was examined (Table 6). Kallisto, while being the fastest, requires a file containing all transcripts sequences. This latest requirement would slow down the pseudoalignment step as all transcripts (ncRNAs, protein-coding, etc.) would have to be entered. A complete list of transcriptome sequences could prove to be heavy and slow down the overall mapping. Furthermore, the list of sequences also impedes the ability to discover new or extended transcripts as only transcripts within the aforementioned list will be taken into account. While exploring the forum dedicated to kallisto, it had been noted that experiments with ncRNAs and small RNA species were lacking. The absence of literature about kallisto's accuracy on smaller species could be problematic as the program was designed to quantify protein coding transcripts. These aspects and the lack of established literature relying on Kallisto dismissed it from further inquiry.

Tophat2, on the other hand, has the ability to discover de novo transcripts. However, as mentioned previously, Tophat2 is based on bowtie2 and while it performs a quicker mapping than its predecessor, the overall time is the highest of all selected programs. Tophat2 also suffers from the poorest positive identification of transcripts, while its memory usage is similar to that of bowtie2. The existence of more accurate and faster performing alternatives to Tophat2 make it superfluous. Tophat2 was not considered as an optimal option for mapping, but STAR offered an alternative. STAR is faster while having a better positive identification of transcripts than Tophat2. As such, STAR was selected to perform mapping of reads to the target genome. Bowtie2 was not used as a first mapping step as it was not splice-aware while the data generated was sure to comprise protein-coding reads.

Examining the output from STAR (Table 5), a large majority of reads, between 80-90%, were properly paired and mapped to the genome. Using a second mapping step with Bowtie2 rescued 5 -10% of reads, the unmapped reads were identified using bowtie2. Bowtie2 was comparable to STAR in terms of runtime, but returned the most positive hits (Table 5). Bowtie2 for its ability to be adjusted to recover smaller reads was used in a second wave of mapping. The overall proper mapping across all datasets were around 85% for the VUS sets while the B*Rz sets were sightly higher on average with >95%. The lower mapping percentage in the VUS sets might come from a multitude of possibilities, however all sets show a high mapping rate to the human genome. The high mapping to the human genome makes it unlikely that the libraries were contaminated.

4.3 Comparison of annotation methodologies

Once the mapping finished, the reads had to be assigned to their corresponding genes. The available and commonly used methods to assign gene annotations were HTSeq, RSEM and bedtools. Bedtools has the lowest performance in term of memory consumption. RSEM is the slowest among the available programs, however it outperforms in its low memory requirements. However, both bedtools and RSEM share a similar disadvantage since both have their main functions in C++. The main functions for normalization of the abundance and peak calling being written in C++ would make it harder to adapt to prediction for ncRNAs and protein-coding transcripts simultaneously. On the other hand, HTSeq is the fastest and has similar CPU requirements to bedtools. HTSeq also has the added advantage of being written entirely in Python and highly customizable. The difference in malleability, but similarity in methods between bedtools and HTSeq made bedtools superfluous for this specific part of the analysis. As for HTSeq and RSEM, a simple method was designed to identify which method performed best with our data. Individual counts returned from each annotation protocol were isolated and compared to the gene's abundances found within the raw bedgraph. The difference between reported and detected accumulation would determine which method is the most accurate (Figure 12A). Both HTSeq and RSEM exhibited erroneous calls. RSEM and HTSeq erroneously attributed most of the reads to the host genes.

However, with the relative ease of modification and customization, HTSeq was adapted to adjust counts and reflect more aptly the data shown by the bedgraph. HTSeq's extra script is referred to as sno_ext and takes advantage of the paired-end nature of the produced NGS data.

The modified version of HTSeq corrects the counts by relying on the overlap between the read and the feature. The features comprise any annotation found within the reference genome. Extra features are also added to the genome annotation by supplementing it with combinations of protein-coding exons. The added protein-coding exons make HTSeq with sno_ext aware of splicing events. Sno_ext addresses splicing by completing the gap between all two consecutive exons in a process dubbed "bridging". This supplemented annotation with the assignment based on overlap ensures that reads falling within exons on both side of ncRNAs encoding intron would not be erroneously labelled as ncRNA. The same principle also corrects overlapping annotations found within host genes (Figure 12B). Often retained introns or lncRNAs overlapping multiple annotations can redirect counts that would otherwise be assigned to ncRNAs to protein-coding or lncRNA transcripts.

4.4 Reads annotation and abundance assessment

Proceeding forward with both versions of HTSeq, all sets were annotated and the gene counts for each was returned to be compared (Table 7). The proportion of reads discarded because of their mapping to multiple locations within the human genome was quite high, varying between 13 and 25%. The higher values found in the VUS datasets were expected since most of the data mapped to snoRNAs, species known for high levels of duplication in humans (Figure 9). HTSeq and the associated modification do not rescue multimapped reads, instead discarding them to remove the possibility of biased accumulation assessment. A second category that is normally disregarded by most analysis methods, the ambiguous reads, the reads that fall within the genomic range of multiple transcripts, are rescued in their entirety with the modification to HTSeq. Sno_ext is able to repatriate ambiguous reads because paired-end sequencing provides precise ends to molecules.

This particularity arises from having a two mates to each read, one for each end of the transcript. By using the exact ends of each read, sno_ext can identify the most likely mapping and provide a valid annotation. The correction affects greatly the overall number of reads annotated in VUS and B*Rz sets, with 30% and close to 10% of reads recovered, respectively. The larger proportion of rescues found in VUS can be explained by the high number of overlaps between snoRNAs and retained introns or lincRNAs. As a matter of fact some of the most expressed snoRNAs, such as SNORD68, are found in host genes with transcripts having retained introns. The lower percentage of ambiguous reads detected in B*Rz sets is associated with the wider range of molecules perceived which are mostly free of overlaps. However, it still remains that without applying the correction some highly abundant snoRNAs are disregarded in further analysis (see figure 12A). As a matter of fact, from the corrected accumulation of BURz_2, 6 out of the 10 most abundant species are snoRNAs while the standard HTSeq analysis returns only 2 snoRNAs from the 6 found within the top 10.

Further inspection of the differences between HTSeq's standard annotation (Figure 8) and that of sno_ext (Figure 9), the most affected sets are the VUS and the BURz. BFRz sets have very little variation whether or not sno_ext modifies the annotation. Both VUS and BURz sets are NGS data coming from sequencing of unfragmented libraries. The most affected species are the same throughout the sets, snoRNAs, protein-coding RNAs and lincRNAs. In the aforementioned sets, the overall abundance of snoRNA species is increased by the correction while both protein-coding RNAs and lincRNAs are substantially lowered. The correction affects snoRNA species counts by rescuing the ambiguous reads, most of which map to snoRNAs, and by reassigning reads erroneously annotated to snoRNAs' host genes.

In the BURz datasets, the rescue of snoRNAs has an additional effect on the detected snoRNAs families. The correction rescues more C/D box snoRNAs than their counterparts because the rescued snoRNAs overlapping with lincRNAs are C/D box snoRNAs. The others did not exhibit the same level between snoRNA families (see figures 6 and 7).

The BFRz remained unaffected by the correction as most of the protein-coding reads had been properly annotated by HTSeq default analysis. VUS, on the other hand, was already preferentially C/D box snoRNAs (~98%). The correction rescues millions of reads, however the overall ratios between H/ACA and C/D box snoRNAs remain almost identical. The species present within VUS sets are different from those found within BURz which accounts for the difference when the correction is applied to the data. Additionally, other datasets generated through the standard protocol relying on the viral retrotranscriptase exhibited an disproportionate representation (>90%) of C/D box snoRNAs such as that found in the VUS sets.

4.5 Comparison between library preparation protocols

Investigating the difference between protocols requires to compared each methods to a gold standard. qPCR was selected as it gives semi-quantitative results and measurements collected within the same experiment can be compared to estimate fold variations between transcripts' abundance. By comparing each library preparation protocol to the qPCR data, it is possible to determine which sequencing workflow reflects more accurately the tissue sample's own populations (Figure 13). The log of the sequencing values was used as to have a linear relationship with the qPCR values. Both Spearman and Pearson coefficients were compared to allow the maximum flexibility in the data interpretation (Table 8). All Spearman coefficients were higher than their Pearson counterparts except for the BURz with the sno_ext correction (Table 8). The difference between both coefficients indicates that the data present in the BURz sets once corrected by sno_ext is the most linear in nature. The other sets exhibit a more monotonic distribution which is more chaotic and difficult to analyze. The VUS sets have low correlation coefficients (Pearson and Spearman) whether or not the data is treated with the sno_ext correction. These low coefficients show that VUS sets do not reflect what is observed by qPCR. On the other hand, the B*Rz sets have higher coefficients values. These values are further improved through sno_ext, often near doubling the coefficients.

These increases aptly demonstrate that the library preparation protocol based on TGIRT offers a better quantification of ncRNAs and that correcting for the oversimplified counting method of HTSeq is required.

4.6 Biases of compositions and size

The analysis does not give a complete overview though, as not all species length and nucleotide compositions are covered by the qPCR analysis. These factors were further assessed by the supplemented ERCC spike-ins B*Rz sets. All injected spike-ins had the identical concentrations which would only be affected by the sequencing workflow. As such, an unbiased sequencing would be expected to have a constant count conserved throughout all spike-ins. However, most NS spike-ins had high fluctuations in their respective counts (Figure 14). All those spike-ins with varying composition were found to be expressed at very low levels with less than a thousand reads associated. Moreover, inspection of the spike-ins with varying length shows a positive linear correlation between length of the spike-in and their detected abundance. The disappearance of the lower molecular weight species came from a purification step to remove primer dimers. Further modifications to the protocol will prevent this phenomenon in a future analysis. This extra purification step aimed at removing dimers might explain the low miRNAs counts found in the B*Rz sets.

5. CONCLUSION

In conclusion, my master's project had two main objectives. The first was to identify a reliable sequencing protocol for the detection and quantification of RNA species, specifically snoRNAs. Examining the available library preparation protocols allowed to remove methods which were known to discard snoRNAs and ncRNAs species. Two remaining options were considered for further analysis, the standard TruSeq small RNA preparation protocol with a step of size selection and the TGIRT protocol. By correlating sequencing data and qPCR data, it was possible to determine that the TGIRT protocol gave more representative accumulation counts. The second goal was to assemble and create a bioinformatics pipeline to detect qualitative and quantitative shifts in the properties of snoRNAs in NGS sequencing datasets. This step was completed through a study of the benchmarking literature and testing. Testing found that mapping with STAR and bowtie2 conserved the most reads when paired in aligning the reads. As for the annotation step, all available methods had flawed read attribution. A customized HTSeq version corrected the issue. The output from the modified HTSeq gave a complete snoRNAs report containing all snoRNA species, their unique sequences, associated counts, the location of each of their boxes as found within the experimental data. While another alternative to HTSeq was found, a portion of the code was brought over to a new pipeline developed in Pr. Michelle Scott's lab.

6. REFERENCES

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Andrews, S. (n.d.). FastQC A Quality Control tool for High Throughput Sequence Data. *Http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*. Retrieved from citeulike-article-id:11583827

Bachellerie, J.-P., Cavaillé, J., & Hüttenhofer, A. (2002). The expanding snoRNA world. *Biochimie*, *84*(8), 775–790. https://doi.org/10.1016/S0300-9084(02)01402-5

Bachellerie, J. P., Cavaillé, J., & Hüttenhofer, A. (2002). The expanding snoRNA world. *Biochimie*, *84*. https://doi.org/10.1016/S0300-9084(02)01402-5

Ballarino, M., Morlando, M., Pagano, F., Fatica, A., & Bozzoni, I. (2005). The Cotranscriptional Assembly of snoRNPs Controls the Biosynthesis of H / ACA snoRNAs in Saccharomyces cerevisiae The Cotranscriptional Assembly of snoRNPs Controls the Biosynthesis of H / ACA snoRNAs in Saccharomyces cerevisiae †. *Molecular and Cellular Biology*, *25(13)*(Juillet), 5396–5403. https://doi.org/10.1128/MCB.25.13.5396

Bizarro, J., Charron, C., Boulon, S., Westman, B., Pradet-Balade, B., Vandermoere, F., … Bertrand, E. (2014). Proteomic and 3D structure analyses highlight the C/D box snoRNP assembly mechanism and its control. *Journal of Cell Biology*, *207*(4), 463–480. https://doi.org/10.1083/jcb.201404160

Blocker, F., Mohr, G., Conlan, L., & Qi, L. (2005). Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. *RNA*, 14–28. https://doi.org/10.1261/rna.7181105.which

Boulon, S., Bertrand, E., & Pradet-Balade, B. (2012). HSP90 and the R2TP co-chaperone complex: Building multi-protein machineries essential for cell growth and gene expression. *RNA Biology*, *9*(2), 148–154. https://doi.org/10.4161/rna.18494

Brameier, M., Herwig, A., Reinhardt, R., Walter, L., & Gruber, J. (2011). Human box C/D snoRNAs with miRNA like functions: Expanding the range of regulatory RNAs. *Nucleic Acids Research*, *39*(2), 675–686. https://doi.org/10.1093/nar/gkq776

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Brosseau, J. P., Lucier, J. F., Lapointe, E., Durand, M., Gendron, D., Gervais-Bird, J., … Elela, S. A. (2010). High-throughput quantification of splicing isoforms. *RNA*, *16*(2), 442–449. https://doi.org/10.1261/rna.1877010

Cavaille, J. (2017). Box C/D small nucleolar RNA genes and the Prader-Willi syndrome: A complex interplay. *Wiley Interdisciplinary Reviews: RNA*. https://doi.org/10.1002/wrna.1417

Cavaillé, J., & Bachellerie, J. P. (1998). SnoRNA-guided ribose methylation of rRNA: Structural features of the guide RNA duplex influencing the extent of the reaction. *Nucleic Acids Research*, *26*(7), 1576–1587. https://doi.org/10.1093/nar/26.7.1576

Chan, P. P., & Lowe, T. M. (2009). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res*, *37*(Database issue), D93-7. https://doi.org/10.1093/nar/gkn787

Chang, H., Lim, J., Ha, M., & Kim, V. N. (2014). TAIL-seq: Genome-wide determination of poly(A) tail length and 3' end modifications. *Molecular Cell*, *53*(6), 1044–1052. https://doi.org/10.1016/j.molcel.2014.02.007

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Deschamps-Francoeur, G., Garneau, D., Dupuis-Sandoval, F., Roy, A., Frappier, M., Catala, M., … Scott, M. S. (2014). Identification of discrete classes of small nucleolar RNA featuring different ends and RNA binding protein dependency. *Nucleic Acids Research*, *42*(15), 10073–10085. https://doi.org/10.1093/nar/gku664

Dieci, G., Preti, M., & Montanini, B. (2009). Eukaryotic snoRNAs: A paradigm for gene expression flexibility. *Genomics*. https://doi.org/10.1016/j.ygeno.2009.05.002

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dupuis-Sandoval, F., Poirier, M., & Scott, M. S. (2015). The emerging landscape of small nucleolar RNAs in cell biology. *Wiley Interdisciplinary Reviews: RNA*, *6*(4), 381–397. https://doi.org/10.1002/wrna.1284

Ender, C., Krek, A., Friedländer, M. R., Beitzinger, M., Weinmann, L., Chen, W., … Meister, G. (2008). A human snoRNA with microRNA-like functions. *Molecular Cell*, *32*(4), 519–28. https://doi.org/10.1016/j.molcel.2008.10.017

Enyeart, P. J., Mohr, G., Ellington, A. D., & Lambowitz, A. M. (2014). Biotechnological applications of mobile group II introns and their reverse transcriptases: gene targeting, RNA-seq, and non-coding RNA analysis. *Mobile DNA*, *5*, 1–19. https://doi.org/10.1186/1759-8753-5-2

Falaleeva, M., & Stamm, S. (2013). Processing of snoRNAs as a new source of regulatory non-coding RNAs: snoRNA fragments form a new class of functional RNAs. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *35*(1), 46–54. https://doi.org/10.1002/bies.201200117

Falaleeva, M., Surface, J., Shen, M., de la Grange, P., & Stamm, S. (2015). SNORD116 and SNORD115 change expression of multiple genes and modify each other's activity. *Gene*, *572*(2), 266–273. https://doi.org/10.1016/j.gene.2015.07.023

Fuller, C. W., Middendorf, L. R., Benner, S. A., Church, G. M., Harris, T., Huang, X., … Vezenov, D. V. (2009). The challenges of sequencing by synthesis. *Nature Biotechnology*, *27*(11), 1013–1023. https://doi.org/10.1038/nbt.1585

Ganot, P., Bortolin, M. L., & Kiss, T. (1997). Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell*, *89*. https://doi.org/10.1016/S0092-8674(00)80263-9

Ganot, P., Caizergues-ferrer, M., & Kiss, T. (1997). The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes & Development*, (11), 941–956.

Ge, J., & Yu, Y. T. (2013). RNA pseudouridylation: New insights into an old modification. *Trends in Biochemical Sciences*. Elsevier Ltd. https://doi.org/10.1016/j.tibs.2013.01.002

Grozdanov, P. N., Fernandez-Fuentes, N., Fiser, A., & Meier, U. T. (2009). Pathogenic NAP57 mutations decrease ribonucleoprotein assembly in dyskeratosis congenita. *Human Molecular Genetics*, *18*(23), 4546–4551. https://doi.org/10.1093/hmg/ddp416

Gumienny, R., Jedlinski, D. J., Schmidt, A., Gypas, F., Martin, G., Vina-Vilaseca, A., & Zavolan, M. (2016). High-throughput identification of C/D box snoRNA targets with CLIP and RiboMeth-seq. *Nucleic Acids Research*, *45*(5), gkw1321. https://doi.org/10.1093/nar/gkw1321

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., … Tuschl, T. (2010). PAR-CliP - A Method to Identify Transcriptome-wide the Binding Sites of

RNA Binding Proteins. *Journal of Visualized Experiments*, (41), 2–6. https://doi.org/10.3791/2034

Havgaard, J. H., Torarinsson, E., & Gorodkin, J. (2007). Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Computational Biology*, *3*(10), 1896–908. https://doi.org/10.1371/journal.pcbi.0030193

Head, S. R., Kiyomi Komori, H., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, *56*(2), 61–77. https://doi.org/10.2144/000114133

Hellemans, J., Mortier, G., De Paepe, A., Speleman, F., & Vandesompele, J. (2007). qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biology*, *8*(2), R19. https://doi.org/10.1186/gb-2007-8-2-r19

Henras, A. K., Dez, C., & Henry, Y. (2004). RNA structure and function in C/D and H/ACA s(no)RNPs. *Curr Opin Struct Biology*, *14*(3), 335–343. https://doi.org/10.1016/j.sbi.2004.05.006

Hoeppner, M. P., & Poole, A. M. (2012). Comparative genomics of eukaryotic small nucleolar RNAs reveals deep evolutionary ancestry amidst ongoing intragenomic mobility. *BMC Evolutionary Biology*, *12*(1), 183. https://doi.org/10.1186/1471-2148-12-183

Holley, C. L., Li, M. W., Scruggs, B. S., Matkovich, S. J., Ory, D. S., & Schaffer, J. E. (2015). Cytosolic accumulation of small nucleolar RNAs (snoRNAs) is dynamically regulated by NADPH oxidase. *Journal of Biological Chemistry*, *290*(18), 11741–11748. https://doi.org/10.1074/jbc.M115.637413

Jorjani, H., Kehr, S., Jedlinski, D. J., Gumienny, R., Hertel, J., Stadler, P. F., … Gruber, A. R. (2016). An updated human snoRNAome. *Nucleic Acids Research*, *44*(11), 5068–5082. https://doi.org/10.1093/nar/gkw386

Kanitz, A., Gypas, F., Gruber, A. J., Gruber, A. R., Martin, G., & Zavolan, M. (2015). Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biology*, *16*(1), 150. https://doi.org/10.1186/s13059-015-0702-5

Kass, S., Tyc, K., Steitz, J. A., & Sollner-Webb, B. (1990). The U3 small nucleolar ribonucleoprotein functions in the first step of preribosomal RNA processing. *Cell*, *60*(6), 897–908. https://doi.org/10.1016/0092-8674(90)90338-F

Kawaji, H., Nakamura, M., Takahashi, Y., Sandelin, A., Katayama, S., Fukuda, S., … Hayashizaki, Y. (2008). Hidden layers of human small RNAs. *BMC Genomics*, *9*(1), 157. https://doi.org/10.1186/1471-2164-9-157

Kehr, S., Bartschat, S., Tafer, H., Stadler, P. F., & Hertel, J. (2014). Matching of soulmates: Coevolution of snoRNAs and their targets. *Molecular Biology and Evolution*, *31*(2), 455–467. https://doi.org/10.1093/molbev/mst209

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res*, *12*. https://doi.org/10.1101/gr.229202. Article published online before March 2002

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. https://doi.org/10.1186/gb-2013-14-4-r36

Kishore, S., Gruber, A. R., Jedlinski, D. J., Syed, A. P., Jorjani, H., & Zavolan, M. (2013). Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome Biol*, *14*(5), R45. https://doi.org/10.1186/gb-2013-14-5-r45

Kishore, S., Khanna, A., Zhang, Z., Hui, J., Balwierz, P. J., Stefan, M., … Stamm, S. (2010). The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. *Human Molecular Genetics*, *19*, 1153–1164. https://doi.org/10.1093/hmg/ddp585

Kishore, S., & Stamm, S. (2006). The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science*, *311*. https://doi.org/10.1126/science.1118265

Kiss-Laszlo, Z., Henry, Y., & Kiss, T. (1998). Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. *EMBO J*, *17*(3), 797–807. https://doi.org/10.1093/emboj/17.3.797

Kiss-László, Z., Henry, Y., & Kiss, T. (1998). Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. *EMBO Journal*, *17*(3), 797–807. https://doi.org/10.1093/emboj/17.3.797

Lambowitz, A. M., & Belfort, M. (2015). Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiology Spectrum*, *3*(2), 1–26. https://doi.org/10.1128/microbiolspec.MDNA3

Lambowitz, A. M., & Zimmerly, S. (2004). Mobile Group II Introns. *Annual Review of Genetics*, *38*(1), 1–35. https://doi.org/10.1146/annurev.genet.38.072902.091600

Langmead. (2013). Bowtie2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923.Fast

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., … Lopez, R. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*. https://doi.org/10.1093/bioinformatics/btm404

Lestrade, L., & Weber, M. J. (2006). snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res*, *34*(Database issue), D158-62. https://doi.org/10.1093/nar/gkj002

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323. https://doi.org/10.1186/1471-2105-12-323

Li, H. (2008). Unveiling substrate RNA binding to H/ACA RNPs: one side fits all. *Current Opinion in Structural Biology*, *18*(1), 78–85. https://doi.org/10.1016/j.sbi.2007.11.004

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, *11*(5), 473–483. https://doi.org/10.1093/bib/bbq015

Li, S., Duan, J., Li, D., Yang, B., Dong, M., & Ye, K. (2011). Reconstitution and structural analysis of the yeast box H/ACA RNA-guided pseudouridine synthase. *Genes Dev*, *25*(22), 2409–2421. https://doi.org/10.1101/gad.175299.111

Li, S., Tighe, S. W., Nicolet, C. M., Grove, D., Levy, S., Farmerie, W., … Mason, C. E. (2014). Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nature Biotechnology*, *32*(9), 915–925. https://doi.org/10.1038/nbt.2972

Locati, M., Terpstra I., de Leeuw W., Kuzak M., Rauwerda H., Ensink W., van Leeuwen S., Nehrdich U., Spaink H., Jonker M., Breit T., and Dekker R. (2015). Improving small RNA-seq by using a synthetic spike-in set for size-range quality control together with a set for data normalization, *Nucleic Acids Research, 43*(14), 89, https://doi.org/10.1093/nar/gkv303

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

McCann, K. L., & Baserga, S. J. (2012). Long noncoding RNAs as sinks in prader-willi syndrome. *Molecular Cell*, *48*(2), 155–157. https://doi.org/10.1016/j.molcel.2012.10.005

McKeegan, K. S., Debieux, C. M., Boulon, S., Bertrand, E., & Watkins, N. J. (2007). A Dynamic Scaffold of Pre-snoRNP Factors Facilitates Human Box C/D snoRNP Assembly. *Molecular and Cellular Biology*, *27*(19), 6782–6793. https://doi.org/10.1128/MCB.01097-07

Michel, C. I., Holley, C. L., Scruggs, B. S., Sidhu, R., Brookheart, R. T., Listenberger, L. L., … Schaffer, J. E. (2011). Small nucleolar RNAs U32a, U33, and U35a are critical mediators of metabolic stress. *Cell Metabolism*, *14*(1), 33–44. https://doi.org/10.1016/j.cmet.2011.04.009

Mohr, S., Ghanem, E., Smith, W., Sheeter, D., Qin, Y., King, O., … Lambowitz, A. M. (2013). Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA (New York, N.Y.)*, 958–970. https://doi.org/10.1261/rna.039743.113

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similiarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nicoloso, M., Qu, L.-H., Michot, B., & Bachellerie, J.-P. (1996). Intron-encoded, Antisense Small Nucleolar RNAs: The Characterization of Nine Novel Species Points to Their Direct Role as Guides for the 2′-O-ribose Methylation of rRNAs. *Journal of Molecular Biology*, *260*(2), 178–195. https://doi.org/10.1006/jmbi.1996.0391

Nottingham, R. M., Wu, D. C., Qin, Y., Yao, J. U. N., Hunicke-smith, S., & Lambowitz, A. M. (2016). RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. *RNA (New York, N.Y.)*, *22*(4), 597–613. https://doi.org/10.1261/rna.055558.115.3

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

O'Neil, D., Glowatz, H., & Schlumpberge, M. (2013). Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Current Protocols in Molecular Biology*. https://doi.org/10.1002/0471142727.mb0419s103

Ono, M., Scott, M. S., Yamada, K., Avolio, F., Barton, G. J., & Lamond, A. I. (2011). Identification of human miRNA precursors that resemble box C/D snoRNAs. *Nucleic Acids Res*, *39*(9), 3879–3891. https://doi.org/10.1093/nar/gkq1355

Peters, J. (2008). Prader-Willi and snoRNAs. *Nat Genet*, *40*(6), 688–689. https://doi.org/10.1038/ng0608-688

Petfalski, E., Dandekar, T., Henry, Y., & Tollervey, D. (1998). Processing of the Precursors to Small Nucleolar RNAs and rRNAs Requires Common Components. *Molecular and Cellular Biology*, *18*(3), 1181–1189. https://doi.org/10.1128/MCB.18.3.1181

Qi, Y., Purtell, L., Fu, M., Lee, N. J., Aepler, J., Zhang, L., … Herzog, H. (2016). Snord116 is critical in the regulation of food intake and body weight. *Scientific Reports*, *6*(1), 18614. https://doi.org/10.1038/srep18614

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Reichow, S. L., Hamma, T., Ferre-D'Amare, A. R., & Varani, G. (2007). The structure and function of small nucleolar ribonucleoproteins. *Nucleic Acids Res*, *35*(5), 1452–1464. https://doi.org/10.1093/nar/gkl1172

Risso, D., Ngai, J., Speed, T. P., & Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples (RUVSeq). *Nature Biotechnology*, *32*(9), 896–902. https://doi.org/10.1038/nbt.2931

Rothé, B., Manival, X., Rolland, N., Charron, C., Senty-Ségault, V., Branlant, C., & Charpentier, B. (2017). Implication of the box C/D snoRNP assembly factor Rsa1p in U3 snoRNP assembly. *Nucleic Acids Research*, 1–19. https://doi.org/10.1093/nar/gkx424

Samarsky, D. A., Fournier, M. J., Singer, R. H., & Bertrand, E. (1998). The snoRNA box C/D motif directs nucleolar targeting and also couples snoRNA synthesis and

localization. *EMBO Journal*, *17*, 3747–3757. https://doi.org/10.1093/emboj/17.13.3747

Schubert, T., & Längst, G. (2013). Changes in higher order structures of chromatin by RNP complexes. *RNA Biology*, *10*(2), 175–179. https://doi.org/10.4161/rna.23175

Schubert, T., Pusch, M. C., Diermeier, S., Benes, V., Kremmer, E., Imhof, A., & Langst, G. (2012). Df31 protein and snoRNAs maintain accessible higher-order structures of chromatin. *Mol Cell*, *48*(3), 434–444. https://doi.org/10.1016/j.molcel.2012.08.021

Scott, M. S., Avolio, F., Ono, M., Lamond, A. I., & Barton, G. J. (2009). Human miRNA precursors with box H/ACA snoRNA features. *PLoS Computational Biology*, *5*. https://doi.org/10.1371/journal.pcbi.1000507

Scott, M. S., & Ono, M. (2011). From snoRNA to miRNA: Dual function regulatory non-coding RNAs. *Biochimie*, *93*(11), 1987–92. https://doi.org/10.1016/j.biochi.2011.05.026

Scott, M. S., Ono, M., Yamada, K., Endo, A., Barton, G. J., & Lamond, A. I. (2012). Human box C/D snoRNA processing conservation across multiple cell types. *Nucleic Acids Res*, *40*(8), 3676–3688. https://doi.org/10.1093/nar/gkr1233

Shendure, J., Ji, H., Huang, H., Wu, C. H., Shendure, J., Ji, H., … Conesa, A. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. https://doi.org/10.1038/nbt1486

Shingara, J. (2005). An optimized isolation and labeling platform for accurate microRNA expression profiling. *RNA*, *11*(9), 1461–1470. https://doi.org/10.1261/rna.2610405

Skryabin, B. V, Gubar, L. V, Seeger, B., Pfeiffer, J., Handel, S., Robeck, T., … Brosius, J. (2007). Deletion of the MBII-85 snoRNA gene cluster in mice results in postnatal growth retardation. *PLoS Genet*, *3*. https://doi.org/10.1371/journal.pgen.0030235

Smith, C. M., & Steitz, J. a. (1997). Sno Storm in the Nucleolus: New Roles for Myriad Small RNPs. *Cell*, *89*(5), 669–672. https://doi.org/10.1016/S0092-8674(00)80247-0

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Molecular Biology*, *147*, 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

System User Guide. (2007), Retrieved from https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-system-guide-15027617-01.pdf.

Taft, R. J., Glazov, E. a, Lassmann, T., Hayashizaki, Y., Carninci, P., & Mattick, J. S. (2009). Small RNAs derived from snoRNAs. *RNA (New York, N.Y.)*, *15*(7), 1233–40. https://doi.org/10.1261/rna.1528909

Trapnell, C., Hendrickson, D., Sauvageau, M., Goff, L., Rinn, J., & Pachter, L. (2012). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology, 31*, 46–53 https://doi.org/10.1038/nbt.2450

Truong, D. M., Sidote, D. J., Russell, R., & Lambowitz, a M. (2013). Enhanced group II intron retrohoming in magnesium-deficient Escherichia coli via selection of mutations in the ribozyme core. *Proc Natl Acad Sci U S A*, *110*(40), E3800-9. https://doi.org/10.1073/pnas.1315742110

Tyc, K., & Steitz, J. A. (1989). U3, U8 and U13 comprise a new class of mammalian snRNPs localized in the cell nucleolus. *The EMBO Journal*, *8*(10), 3113–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=401391&tool=pmcentrez&rendertype=abstract

Tycowski, K. T., Shu, M. D., & Steitz, J. A. (1993). A small nucleolar RNA is processed from an intron of the human gene encoding ribosomal protein S3. *Genes and Development*, *7*(7 A), 1176–1190. https://doi.org/10.1101/gad.7.7a.1176

Tycowski, K. T., Smith, C. M., Shu, M. D., & Steitz, J. A. (1996). A small nucleolar RNA requirement for site-specific ribose methylation of rRNA in Xenopus. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(25), 14480–5. https://doi.org/10.1073/pnas.93.25.14480

Walbott, H., Machado-pinilla, R., Liger, D., Grozdanov, P. N., Godin, K., Tilbeurgh, H. Van, … Meier, U. T. (2011). The H / ACA RNP assembly factor SHQ1 functions as an RNA mimic The H / ACA RNP assembly factor SHQ1 functions as an RNA mimic, 2398–2408. https://doi.org/10.1101/gad.176834.111

Watkins, N. J., Dickmanns, A., & Luhrmann, R. (2002). Conserved stem II of the box C/D motif is essential for nucleolar localization and is required, along with the 15.5K protein, for the hierarchical assembly of the box C/D snoRNP. *Mol Cell Biol*, *22*(23), 8342–8352. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12417735

Watkins, N. J., Segault, V., Charpentier, B., Nottrott, S., Fabrizio, P., Bachi, A., … Luhrmann, R. (2000). A common core RNP structure shared between the small nucleoar box C/D RNPs and the spliceosomal U4 snRNP. *Cell*, *103*(3), 457–466. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11081632

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., … Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Research*, *44*(D1), D710–D716. https://doi.org/10.1093/nar/gkv1157

Yin, Q. F., Yang, L., Zhang, Y., Xiang, J. F., Wu, Y. W., Carmichael, G. G., & Chen, L. L. (2012). Long noncoding RNAs with snoRNA ends. *Mol Cell*, *48*(2), 219–230. https://doi.org/10.1016/j.molcel.2012.07.033

Youssef, O. A., Safran, S. A., Nakamura, T., Nix, D. A., Hotamisligil, G. S., & Bass, B. L. (2015). Potential role for snoRNAs in PKR activation during metabolic stress. *Proceedings of the National Academy of Sciences*, *112*(16), 5023–5028. https://doi.org/10.1073/pnas.1424044112

Zhang, X.-O., Yin, Q.-F., Wang, H.-B., Zhang, Y., Chen, T., Zheng, P., … Yang, L. (2014). Species-specific alternative splicing leads to unique expression of sno-lncRNAs. *BMC Genomics*, *15*(1), 287. https://doi.org/10.1186/1471-2164-15-287

Zhao, W., He, X., Hoadley, K. A., Parker, J. S., Hayes, D., & Perou, C. M. (2014). Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics*, *15*(1), 419. https://doi.org/10.1186/1471-2164-15-419

Zheng, G., Qin, Y., Clark, W. C., Dai, Q., Yi, C., He, C., … Pan, T. (2015). Efficient and quantitative high-throughput tRNA sequencing. *Nature Methods*, *12*(9), 835–7. https://doi.org/10.1038/nmeth.3478